

Mining User-Generated Content for Incidents

Auswertung von nutzergenerierten Inhalten für Schadensereignisse

Zur Erlangung des akademischen Grades Doktor-Ingenieur (Dr.-Ing.)

genehmigte Dissertation von Dipl.-Wirtsch.-Inform. Axel Schulz aus Magdeburg

Mai 2014 — Darmstadt — D 17



Telecooperation Lab



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Department of Computer Science
Telecooperation

Mining User-Generated Content for Incidents
Auswertung von nutzergenerierten Inhalten für Schadensereignisse

Genehmigte Dissertation von Dipl.-Wirtsch.-Inform. Axel Schulz aus Magdeburg

1. Gutachten: Prof. Dr. Max Mühlhäuser
2. Gutachten: Prof. Dr. York Sure-Vetter

Tag der Einreichung: 01.05.2014

Tag der Prüfung: 23.06.2014

Darmstadt — D 17

"To leave the world a little better than you found it. That's the best a man can ever do."
- Paul Auster



Erklärung zur Dissertation

Hiermit versichere ich, die vorliegende Dissertation ohne Hilfe Dritter nur mit den angegebenen Quellen und Hilfsmitteln angefertigt zu haben. Alle Stellen, die aus Quellen entnommen wurden, sind als solche kenntlich gemacht. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Darmstadt, den 01. Mai 2014

(Axel Schulz)



Abstract

Social media changed human interaction by allowing people to connect to each other anytime and from anywhere, resulting in many valuable information shared about a variety of different domains. Because of the large amount of data created every day, social media analytics became an important topic to make use of this source of information. Most importantly, people contribute much valuable information about crisis events such as small-scale incidents, which is currently not taken into account by decision makers in emergency management. Reasons for this are the sheer amount of information as well as the heterogeneous and unstructured nature of the data, which hinder the use of this source of information.

In this dissertation, we try to answer the question *"How can user-generated content be made a usable and valuable source of information for situational awareness of decision makers?"*. To answer this question, we present a framework consisting of the necessary steps to process a large amount of social media data in such a way that incident-related information is identified and aggregated to distinct incident clusters, each one containing information about an individual real-world incident. With the contributions presented in this dissertation, previously not usable user-generated content becomes a valuable source for decision making in emergency management.

In the first part of the dissertation, we introduce the requirements of a system for small-scale incident detection, which are derived from the (1) spatial, (2) temporal, and (3) thematic dimensions defining an incident. Based on these dimensions, we develop a framework for processing a large amount of user-generated content. As a first step of the framework, we introduce how user-generated content is collected to create an initial information base, which is processed in the subsequent steps of the framework.

In the second part, we introduce several steps to preprocess unstructured social media data so it can be used in the subsequent steps of the framework. We show how named entities and temporal expressions are identified and extracted so that they can be used as additional information when creating a machine learning model that is generalizable for data that stems from a different city. We introduce a set of adaptations applied to standard techniques that allow us to extract named entities and temporal expressions from unstructured text. We also present how we make use of the temporal expressions to infer the point in time when an incident occurred.

In that part, we also deal with the problem of how to infer the spatial dimensions of user-generated content. We contribute a novel approach for the geolocalization of tweets that is capable of inferring the home location of a Twitter user, the point of origin where a tweet was sent, as well as of inferring the location focus of a

tweet message. We show that the approach is able to locate 92% of all tweets with a median accuracy of below 30 km, thus outperforming related approaches. Furthermore, it predicts the user's residence with a median accuracy of below 5.1 km. Finally, the same approach is able to estimate the focus of incident-related tweets within a median accuracy of below 250 m.

In the third part of the dissertation, we present approaches for inferring the thematic dimension. We contribute a general approach for applying crowdsourcing to *manually* classify and aggregate user-generated content according to the information need of the command staff in emergency management. With this approach, we are able to differentiate incident-related information from information not related to an incident. Evaluation results with end users show that this approach is indeed valuable for the command staff.

As crowdsourcing is limited when it comes to the timely filtering of a large amount of information, we further contribute an approach for *automatically* detecting incident-related information in user-generated content. Based on an extensive evaluation of different feature groups, we present a highly precise machine learning approach that is able to classify the incident type with an F-measure of more than 90%. We also deal with the dynamism and regional variation of user-generated content and contribute a concept that allows creating features that are not city-specific and support training a generalized model.

Based on the temporal, spatial, and thematic dimensions of each information item, we present a clustering approach that is able to detect incidents in a large amount of social media data. The approach clusters all information related to the same incident. Furthermore, it is able to cope with different organizational incident type vocabularies. Evaluation results show that the approach is able to detect more than 50% of real-world incidents published in an official emergency management system. Furthermore, 32.14% of the detected incidents are within a 500 m radius and within a 10 min time interval of the real-world incident, allowing precise spatial and temporal localization. With this recall, we outperform related approaches, which only detect about 5% of the real-world incidents. Also, more than 77% of the incident clusters created with our approach are indeed related to incidents, thus significantly reducing the quantity of irrelevant information. Furthermore, we underline the importance of incident-related tweets by conducting a user study of situational information shared in user-generated content, showing that valuable situational information is indeed shared in social media.

Finally, we contribute an approach for reducing labeling costs of user-generated content. The presented algorithms make use of temporal, spatial, and thematic metadata to determine the most valuable instances to label. Our evaluation shows that the approach outperforms current state-of-the-art approaches. Furthermore, we show that labeling costs can indeed be reduced.

Zusammenfassung

Soziale Netzwerke, wie z.B. Twitter, ermöglichen es, sich von jedem Ort und zu jeder Zeit miteinander zu verbinden. Dies führt dazu, dass eine große Menge an Informationen über vielfältige Themen erzeugt wird. Ein Teil dieser nutzergenerierten Inhalte bezieht sich auf alltägliche Ereignisse, wie Autounfälle oder Brände. Obwohl diese Informationen zu Schadensereignissen potentiell nützlich für die Entscheidungsfindung im Krisenmanagement sein können, besteht momentan keine Möglichkeit diese zu verwenden. Dies liegt zum einen daran, dass eine manuelle und zeitnahe Verarbeitung der Menge an Informationen nur schwer durchführbar ist. Zum anderen ist eine automatisierte Verarbeitung dieser größtenteils unstrukturierten Daten bisher nicht möglich.

In der vorliegenden Dissertation soll die Frage behandelt werden, wie nutzergenerierte Inhalte aufbereitet werden können, um diese zu einer nützlichen und wertvollen Informationsquelle für Entscheidungsträger zu machen. Dazu wird in dieser Arbeit ein System vorgestellt, welches die notwendigen Verarbeitungsschritte durchführt, um Informationen, die sich auf Schadensereignisse beziehen, zu erkennen und zu aggregieren.

Im ersten Teil der Dissertation werden zunächst die Anforderungen an ein System vorgestellt, welches Schadensereignisse auf Basis von nutzergenerierten Inhalten erkennt. Diese Anforderungen werden von drei Eigenschaften, die ein Schadensereignis definieren, abgeleitet: der *räumlichen*, *zeitlichen* und *thematischen* Dimension. Basierend auf diesen Dimensionen wurde ein System entwickelt, um Schadensereignisse auf Basis großer Datenmengen zu erkennen.

Als erster Schritt des Systems wird beschrieben, wie eine initiale Informationsbasis aus sozialen Medien geschaffen wird. Als zweiter Schritt werden unterschiedliche Verfahren präsentiert, um nutzergenerierte Inhalte derart aufzubereiten, dass diese in den darauffolgenden Verarbeitungsschritten genutzt werden können. Dazu wird zunächst gezeigt, wie Eigennamen und zeitliche Ausdrücke erkannt und extrahiert werden können, um diese als zusätzliche Informationen für die Bestimmung der drei Dimensionen zu nutzen. Weiterhin werden zeitliche Ausdrücke dafür eingesetzt, um den genauen Zeitpunkt zu bestimmen, wann ein Schadensereignis stattgefunden hat.

In diesem Teil der Dissertation wird ebenfalls das Problem adressiert, wie die räumliche Dimension eines Schadensereignisses ermittelt werden kann, d.h. wie nutzergenerierte Inhalte räumlich verortet werden können. Dafür wird ein Algorithmus vorgestellt, welcher in der Lage ist, den Heimatort eines Twitternutzers, seinen Standort zum Zeitpunkt des Sendens einer Nachricht sowie den Bezugspunkt der Nachricht

zu ermitteln. Es wird gezeigt, dass dieser Ansatz in der Lage ist, 92% aller Tweets in einem Median von unter 30 km zu verorten. Weiterhin ist der Ansatz in der Lage den Heimatort mit einem Median unter 5.1 km sowie den Bezugspunkt einer Nachricht mit einem Median unter 250 m zu bestimmen. Mit dieser Genauigkeit werden die Ergebnisse bisheriger Ansätze übertroffen.

Im dritten Teil der Dissertation werden zunächst Ansätze präsentiert, um die thematische Dimension einer Information zu bestimmen. Es wird gezeigt, wie Crowdsourcing eingesetzt werden kann, um nutzergenerierte Inhalte *manuell* zu klassifizieren und zu aggregieren, so dass die relevanten Informationen für die Entscheidungsfindung eines Entscheidungsträgers gefiltert werden können. In einer Evaluation mit Entscheidungsträgern wurde gezeigt, dass dieser Ansatz in der tagtäglichen Entscheidungsfindung nützlich ist.

Da Crowdsourcing nur begrenzt einsetzbar ist, wenn zeitnah große Datenmengen gefiltert werden müssen, wird weiterhin ein Ansatz zur *automatischen* Klassifikation der thematischen Dimension beschrieben. Dafür wird ein präzises Modell zum maschinellen Lernen vorgestellt, um den Typ des Schadensereignisses mit einer Genauigkeit von mehr als 90% zu klassifizieren. Weiterhin wird gezeigt, dass dieses Modell robust gegenüber regionalen Unterschieden in der Verwendung von Begriffen, wie z.B. der Verwendung von lokalen Straßennamen, ist.

Basierend auf den in den einzelnen Ansätzen ermittelten räumlichen, zeitlichen und thematischen Informationen, wird ein Ansatz zum Aggregieren zusammengehöriger Informationen beschrieben. In einer Evaluation auf Basis von Daten aus einem Krisenmanagementsystem wird gezeigt, dass dieser Ansatz in der Lage ist 50% der echten Schadensereignisse zu erkennen, sowie dass mehr als 32% der erkannten Ereignisse innerhalb eines 500 m Radius und innerhalb eines 10 min Intervalls liegen. Mit diesem Ansatz wird daher eine präzise räumliche und zeitliche Verortung ermöglicht, die deutlich bestehende Ansätze übertrifft. Weiterhin wird gezeigt, dass 77% der erkannten Ereignisse tatsächlich Schadensereignisse sind, wodurch die Menge an irrelevanten Informationen deutlich verringert werden kann.

Da die automatische Klassifikation der thematischen Dimension eine manuelle Klassifikation voraussetzt, wird abschließend ein Ansatz beschrieben, um die Kosten dieser manuellen Klassifikation zu senken. Dazu werden zwei Algorithmen zur Ereignisbasiertes Aggregation vorgestellt, welche bestmögliche Instanzen zum manuellen Klassifizieren auswählen. In einer Evaluation wird gezeigt, dass dieser Ansatz bessere Ergebnisse erzielt als bisherige Ansätze.

Mit den in dieser Arbeit erbrachten Beiträgen ist es letztlich möglich, bisher nicht nutzbare Informationen aus sozialen Medien für die Entscheidungsfindung im Krisenmanagement nutzbar zu machen.

Acknowledgments

This thesis would not have been possible without the support of my advisors, colleagues, students, family, and friends.

In particular, I thank my professor and advisor Prof. Dr. Max Mühlhäuser (Technische Universität Darmstadt, Germany) for enabling me to do this dissertation. I am very thankful for his continuous support during the last three years. Furthermore, I thank Prof. Dr. York Sure-Vetter for being my co-referee.

Special thanks go to Florian Probst (SAP AG) and Immanuel Schweizer (TU Darmstadt) for mentoring and advising me during this time. Thanks Florian, for many ontological discussions and never-ending one pager writing, and thank you, Immanuel, for keeping the heavy workload away from me and taking care of my funding. Also, thanks go to my manager, Knut Manske, for helping me to focus on my dissertation at SAP.

Among the different colleagues that supported me throughout these years, I particularly thank Daniel Bär, Markus Döhring, Christian Guckelsberger, Frederik Janssen, Kamill Panitzek, Heiko Paulheim, Benedikt Schmidt, Sebastian Döweling, and Thorsten Strufe. Also, I like to thank all of my students who contributed to my research projects, in particular, Max Bruchmann, Jakob Karolus, Felix Mayer, Johannes Nachtwey, Petar Ristoski, and Tung Dang Thanh. Further, I am immensely thankful for all the support I received from present and former members of the Telecooperation Lab at Technische Universität Darmstadt and the SAP Research lab in Darmstadt. Working with all of you inspired my work and it was a pleasure.

Last but not least, I am deeply grateful for the support of my beloved wife, Jasmin, my family, and my parents-in-law.



Contents

1. Introduction	1
1.1. Motivation	1
1.2. Research Questions	2
1.3. Challenges	3
1.4. Research Scope	4
1.5. Contributions and Outline	5
I. Initial Foundations for Detecting and Clustering Incident-Related Information in User-Generated Content	9
2. Design Considerations of a Framework for Detecting Incidents	11
2.1. Initial Definitions and Design Considerations	11
2.2. Architecture of a Framework for Small-Scale Incident Detection	13
2.3. Conclusion	16
3. Considerations of Collecting and Filtering User-Generated Content	17
3.1. Background on User-Generated Content and Twitter	17
3.2. Data Collection and Filtering of User-Generated Content	19
3.2.1. Data Sets	19
3.2.2. Incident Types and Keyword Filtering	20
3.3. Conclusion	22
II. Automatic Preprocessing and Geolocalization of User-Generated Content	23
4. Preprocessing of Unstructured Text	27
4.1. Natural Language Preprocessing	28
4.2. Named Entity and Temporal Expression Recognition on Unstructured Text	30
4.2.1. Definition of Named Entities and Temporal Expressions	31
4.2.2. Named Entity Recognition and Replacement Using Linked Open Data	32
4.2.3. Location Mention Extraction and Replacement	33
4.2.4. Temporal Expression Recognition and Normalization on Unstructured Text	34

4.2.5. Evaluation of Named Entity and Temporal Expression Recognition	36
4.3. Conclusion	40
5. Geolocalization of User-Generated Content	41
5.1. Background	42
5.1.1. Toponym Resolution	43
5.1.2. Spatial Indicators in Tweets	43
5.2. Related Work	46
5.2.1. The Use of Spatial Indicators in Related Work	46
5.2.2. Techniques Used in Related Work	49
5.2.3. Focus of Geolocalization in Related Work	49
5.2.4. Discussion of Related Work	49
5.3. Approach	50
5.3.1. Approach for Determining the Place of Origin of a Tweet and the Home Location of a Twitter User	50
5.3.2. Approach for Estimating the Focus of Incident-Related Tweets	56
5.4. Evaluation	58
5.4.1. Data Set and Metrics	59
5.4.2. Determining External Quality Measures	61
5.4.3. Study 1: Evaluation of Single Spatial Indicators	61
5.4.4. Study 2: Evaluation of Estimating the Place of Origin	66
5.4.5. Study 3: Evaluation of Estimating the Home Location	67
5.4.6. Study 4: Evaluation of Estimating the Focus of Incident-Related Tweets	69
5.5. Conclusion	70

III. Incident Detection and Clustering of Incident-Related Information 73

6. Human-Based Classification and Aggregation of User-Generated Content	79
6.1. Background	80
6.2. Related Work	83
6.3. Approach	89
6.4. Prototypical Implementation	94
6.5. Evaluation	94
6.5.1. Study 1: Evaluation of Human-Centered Sensing and Human-Based Classification in Emergency Management	96
6.5.2. Study 2: Qualitative Evaluation of Human-based Classifications	100
6.6. Conclusion	104

7. Machine-Based Classification of User-Generated Content	107
7.1. Background	109
7.2. Related Work	112
7.2.1. Incident Type Classification	113
7.2.2. Approaches Training a Generalized Model For Classifying User-Generated Content	116
7.2.3. Summary	118
7.3. Approach	118
7.3.1. Semantic Abstraction	121
7.3.2. Feature Generation	124
7.3.3. Classification	128
7.4. Prototypical Realization	128
7.5. Evaluation	129
7.5.1. Data Sets, Metrics, and Methodology	129
7.5.2. Study 1: Incident Type Classification Based on Keywords	133
7.5.3. Study 2: Incident Type Classification - Initial Feature Selection	134
7.5.4. Study 3: Incident Type Classification - Semantic Abstraction . .	142
7.5.5. Study 4: Evaluation of Generalizability Using Semantic Ab- straction	151
7.6. Conclusion	166
8. Machine-Based Aggregation of User-Generated Content	169
8.1. Related Work	171
8.1.1. Type of Event	175
8.1.2. Clustering Approach	179
8.1.3. Metadata Used	181
8.1.4. Summary	182
8.2. Approach	183
8.2.1. Incident Detection and Clustering of Related Information	183
8.2.2. Clustering Across Different Vocabularies	188
8.2.3. Aggregation of Incident Reports to Incident Clusters	189
8.3. Prototypical Realization	189
8.4. Evaluation	193
8.4.1. Study 1: Analysis of Incident Reports	193
8.4.2. Study 2: Evaluation of Incident Detection	201
8.5. Conclusion	207
9. Refinement of the Framework for Detecting and Clustering Incident In- formation	209
9.1. Background	211
9.2. Related Work	214
9.2.1. Active Learning on Social Media Data	214

9.2.2. Selection Strategies for Active Learning	215
9.2.3. Active Learning and Noisy Labels	216
9.2.4. Summary	216
9.3. Approach	217
9.3.1. Motivating Example	218
9.3.2. Event-Based Clustering for Active Learning	220
9.3.3. Initial Selection Strategy	221
9.3.4. Query Selection Strategy	222
9.4. Evaluation	224
9.4.1. Experimental Setup	225
9.4.2. Results	228
9.4.3. Summary	235
9.5. Conclusion	235
10. Conclusions	237
10.1. Summary	237
10.2. Future Research Directions	239
A. Appendix	243
A.1. Evaluation of Semantic Abstraction in Addition to the Best Feature Groups	243
A.2. Evaluation Results of Machine-Based Aggregation of User-Generated Content	247
Bibliography	249
List of Figures	275
List of Tables	279
List of Algorithms	283

1 Introduction

Social media changed the way of human interaction by allowing people to connect to each other anytime and from anywhere. Knowledge and information can be shared quickly on a variety of platforms. Because of the large amount of data created every day, social media analytics became an important topic to make use of this source of information. For instance, social media is used to share opinions regarding movies or products, which is valuable for market research [16, 189]. Political orientation [52, 53, 242] and medical conditions [223, 204] are expressed. Furthermore, information about events such as music festivals or soccer games is shared [167].

Recent research showed the relevance of social media for emergency management. With the increasing adoption of smartphones with multiple sensors and a constant Internet connection [35], humans act as *social sensors* and become a valuable source of timely incident information. For instance, valuable information was shared during incidents like the Oklahoma grass fires and the Red River floods in April 2009 [236, 219] or the terrorist attacks in Mumbai [85, 164]. When the US Airways flight 1549 crashed into the Hudson River, four minutes after the crash, the first message was shared on Twitter, 15 minutes before the mainstream media [20]. Also, after a storm hit the Pukkelpop festival in Belgium in 2011, bystanders shared important situational information [227]. Furthermore, the analysis of social media showed its value for detecting earthquakes in real time or tracking diseases [192, 213].

In summary, user-generated content gained a steadily increasing relevance in different areas. Social media is a rich source of incident-related information, especially for emergency management.

1.1 Motivation

Improving situational awareness (i.e., "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" [68]) is one of the main goals for efficient decision making in emergency management. Situational awareness is crucial for the success of the operation in the first 72 hours of a disaster [72].

At the beginning of a crisis, the emergency management staff begins to gather and analyze information from multiple sources. On the one side, current information provided by on-site rescue squads and information from internal emergency management systems are used. On the other side, with the changing information landscape in the World Wide Web, citizens submit their observations and inferences about real-world events more frequently using social media. Different situational information

that could contribute to detect incidents or that would help to understand the situation at hand is already present. This information might help to improve situational awareness.

Decision makers in crisis management could highly benefit from this new source of information, if appropriate, and reliable information from citizens could be retrieved in time. However, this information source remains unused. One reason for this is the sheer amount of information created every day, which results in an *information overload* that is not manageable for humans. Also, automatic processing is not applied, because of the heterogeneous and unstructured nature of the data. Thus, inferences on all available information are not easily drawn and potentially valuable situational information remains unused.

Nevertheless, it is unquestioned that the massive stream of user-generated content contains pieces of highly relevant information that is not known to the decision maker. Harvesting this information can contribute to a better situational picture finally leading to an improved situational awareness compared with a situation where this information is not available at all. But up to now, no reliable approach exists for retrieving high-quality information in the vast amount of available user-generated content.

This results in the key question to answer in this dissertation:

How can user-generated content be made a usable and valuable source of information for situational awareness of decision makers?

To answer this question, we follow the vision of a framework that supports a decision maker by providing inferences about an incident based on information present in social media. These inferences can be used to increase situational awareness, leading to more informed decisions, to shorter reaction times, and, finally, to a higher number of successful operations.

To realize this vision, several subordinated questions have to be answered (see Section 1.2) with respect to certain challenges (see Section 1.3). The scope for this dissertation is set in Section 1.4. Finally, the structure and the contributions of this dissertation are provided in Section 1.5.

1.2 Research Questions

The goal of this dissertation is to answer *How can user-generated content be made a usable and valuable source of information for situational awareness of decision makers?*. To answer this question, we aim to detect incident-related information in the vast amount of user-generated content. Thus, our goal is to develop a framework that enables (1) *the detection of incidents based on user-generated content* and (2) *the aggregation of incident-related information* in such a way that previously unused user-

generated content becomes a usable and valuable source of information for decision makers.

To achieve this, we have to tackle the following subordinated questions:

- *How can incident-related information be manually and automatically detected in user-generated content?*

One of the major issues is to identify incident-related information in the vast amount of user-generated content. For manual filtering, crowdsourcing may be applied, whereas data mining is suitable for automatic processing. In the first case, it is important to understand how to apply crowdsourcing on user-generated content. In the latter case, it needs to be investigated if it is possible to build an automatic classification approach that is able to classify user-generated content with high performance and accuracy. Furthermore, it needs to be understood which preprocessing steps are needed to convert unstructured information to structured information to enable automatic processing.

- *How can information related to the same incident be manually and automatically aggregate?*

As different information might refer to the same incident, a clustering of this information according to the incident it refers to is necessary. Thus, considering the impact of time and space is important as each incident has a fixed temporal and spatial extent. For this, it is essential to understand how to infer temporal and spatial information from user-generated content. Furthermore, new means are needed for aggregating related information to provide it in a way it can be used in emergency management.

1.3 Challenges

Social media data is rather different compared with other information sources. It has different characteristics that complicate answering the research questions. These characteristics are described in the following:

- **Vast amount of information:** The amount of social media data created every day is further increasing [6]. This results in an information overflow, which is difficult to handle. There is a lack of time to analyze the incoming flood of data, especially for time critical decisions.
- **Heterogeneity:** The types of social media data differ. Social media content might be audio or video files, images, or textual content. This content is not necessarily interlinked. Furthermore, it is shared across various platforms.
- **Dynamism:** Information in social media is changing frequently. For instance, people update their current location or their current status. Furthermore, inter-

ests change rapidly as trends evolve. Thus, user-generated content has a very dynamic nature.

- **Reliability:** Social media platforms are used by companies, domain experts, as well as a variety of regular users. Also, these platforms are spammed by automatic bots and people alike. This results in a high variety of quality, which makes the identification of relevant and reliable information much harder.
- **Interconnectedness:** Compared with traditional texts, textual data in social media is not independent and identically distributed (i.i.d.) [6]. For instance, people annotate their content with specific annotations such as hashtags, which are used to refer to a certain topic. Also, users share URLs that refer to external websites. Furthermore, users themselves are interlinked with each other via friendship or follower relationships.

In particular, textual content shared in social media has special properties that pose new challenges to our research goal:

- **Unstructuredness of textual content:** Text shared in social media is inherently unstructured. Users tend to use abbreviations or nonstandard vocabulary in their posted content. This is even increased through the diversity of authorship; thus, many different styles of writing can be found. Some users such as domain experts post information carefully, while other users do not.
- **Length of textual content:** In most social networks, the length of each posting is limited. For instance, messages on Twitter are limited to 140 characters. Thus, short messages consist of only few phrases or sentences.
- **Regional variation:** Words and phrases used in social media texts are interconnected to the location where a text was created. Thus, mechanisms that apply for one city may not necessarily apply just as precisely for data of a different city.

1.4 Research Scope

The scope of this dissertation is constrained by the following aspects:

- **Small-scale incidents:** Within this dissertation, only small-scale incidents are considered. Detecting small-scale incidents in the vast amount of user-generated content poses new challenges to algorithms in this area. The absolute amount of information available for an everyday small-scale event is low. Thus, the detection, analysis, and usage of information related to such small-scale incidents are much harder. During large-scale events, postings are usually plentiful, and missing a small-amount might be irrelevant. But with

only a small amount of postings available, this can make a world of difference for decision making in emergency management.

- Pool-based sampling: In this dissertation, we follow a pool-based approach [134] as we assume that all data used is already present for analysis. As a large amount of social media data can be collected in seconds, this is a valid restriction. However, approaches that are applicable in a pool-based setting can also be reused in a stream-based approach.
- Textual content: In this dissertation, we focus on textual content as there has been a rapid growth of text data in social media [6]. The approaches presented in this dissertation are general approaches; however, we mainly rely on data retrieved from the social network Twitter. Twitter was used as one very prominent platform on which information is shared every day and by a variety of people. In 2013, Twitter had about 240 million active users [232], who shared more than 500 million messages per day [205]. This huge amount of data gives a wide base of information for a variety of topics. We decided to use tweets as one frequently updated source of user-generated content, which is easy to collect. Furthermore, it has been shown that tweets are created in real time and shortly after an incident occurred; thus, they provide a valuable source of incident-related information. Additionally, Twitter provides rich metadata such as a time stamp or even GPS coordinates of a user. We expect that approaches that perform sufficiently on tweets also perform well on other social media texts.

1.5 Contributions and Outline

This dissertation is separated into three parts:

- Part I - A Framework for Detecting and Clustering Incident Information in User-Generated Content
 - In Chapter 2, we introduce the definition of an event and an incident as a specific type of event. We show that an event can clearly be characterized by a temporal, spatial, and thematic dimension. Based on these definitions, we identify requirements of a system that is able to identify incident-related information in user-generated content. Based on these requirements, we present the general architecture of our framework that consists of the necessary steps to help us answer the question *How can user-generated content be made a usable and valuable source of information for decision makers?*.
 - As a first step of the framework, user-generated content is collected. In Chapter 3, we describe how an initial information base is created, which

can further be processed in the subsequent steps of the framework. For this, we present background on user-generated content and Twitter as our major source of incident-related information. Furthermore, we give an overview of the data collection setup.

- Part II - Automatic Preprocessing of User-Generated Content
 - As the texts shared in social media are unstructured, further processing is needed. Chapter 4 provides preliminaries needed for preparing text so it can be used for the subsequent steps of the framework. In the same chapter, we show how named entities and temporal expressions are identified and extracted so that they can be used as additional information when applying automatic processing techniques. We present a set of adaptations applied to standard techniques that allow us to extract named entities and temporal expressions from unstructured text. We also present how we make use of the temporal expressions to infer the point in time when an incident occurred.
 - In the second chapter of this part (see Chapter 5), we deal with the problem how to infer spatial information from user-generated content. For this, we identify parts of tweets and their metadata suitable for geolocalization. In that chapter, we present the first major contribution of this dissertation. We propose a novel approach for the geolocalization of tweets that is capable of inferring the home location of a Twitter user, the point of origin where a tweet was sent, as well as for inferring the location focus of a tweet message. We validate the accuracy of our approach and show that the approach is able to locate 92% of all tweets with a median accuracy of below 30 km. Furthermore, it predicts the user's residence with a median accuracy of below 5.1 km. Finally, the same approach is able to estimate the focus of incident-related tweets within a median accuracy of below 250 m.

The contributions are partly published in [201, 200].

- Part III - Incident Detection and Clustering of Incident-Related Information
 - In Chapter 6, we contribute a general approach for applying crowdsourcing to classify and aggregate user-generated content according to the information need of the command staff in emergency management. With this approach, we are able to *manually* differentiate incident-related information from information not related to an incident. Also in this step, we present human-centered sensing as a means for collecting additional information about an incident. Our evaluation shows that these approaches are indeed valuable for the command staff.

-
- As crowdsourcing is limited when it comes to the timely filtering of a large amount of information, we present an approach for *automatically* detecting incident-related information in user-generated content (see Chapter 7). For this, we present an extensive evaluation for determining optimal feature sets for this task. We validate the performance of the best feature combination on different data sets and show that we are able to classify the incident type with an F-measure of more than 90%. Also, we deal with the dynamism and regional variation of user-generated content. For this, we introduce the novel concept of semantic abstraction, which allows the creation of features that are not city-specific and support training a generalized model. We evaluate semantic abstraction on data sets from five different cities and show that it is indeed a valuable means.
 - Based on the information inferred in the preceding chapters, we contribute a spatio-temporal-thematic clustering approach, which is able to detect incidents in a large amount of social media data (see Chapter 8). The approach clusters all information related to the same incident and is able to deal with different organizational incident type vocabularies. We evaluate the approach and show that we are able to detect more than 50% of real-world incidents published in an official emergency management system. Furthermore, 32.14% of the detected incidents are within a 500 m radius and within a 10 min time interval of the real-world incident, allowing precise spatial and temporal localization. These results are more than five times better compared with related approaches. Also, more than 77% of the incident clusters created with our approach are indeed related to incident events. Furthermore, we evaluate the value of situational information shared in tweets posted in two North American cities. We show that a variety of individuals share information about small-scale incidents. Furthermore, we show that important situational information about affected objects, injured persons, and the location of an incident is shared, which is important information for decision making.
 - As the last chapter in this part, we present an approach for refining the framework according to different information needs (see Chapter 9). This is important as the machine-based approaches need to be adapted to changing conditions such as different incident types or different information sources. Furthermore, we deal with the problem that refining the framework is costly as new information needs to be collected with extensive human effort. To manage these aspects, we present a novel event-based clustering approach that makes use of spatial, temporal, and thematic information. We validate the effectiveness of our approach on a data set of incident-related tweets compared with state-of-the-art approaches and show that our approach outperforms related work. Further-

more, the refinement step helps to reduce the amount of information that needs to be processed manually, thus reducing the overall costs for refining the framework.

The contributions are partly published in [194, 196, 197, 201, 195, 202, 203, 199, 198, 154]. Also, one approach is patented [88].

In Figure 1, an overview of the connections between all chapters is shown.

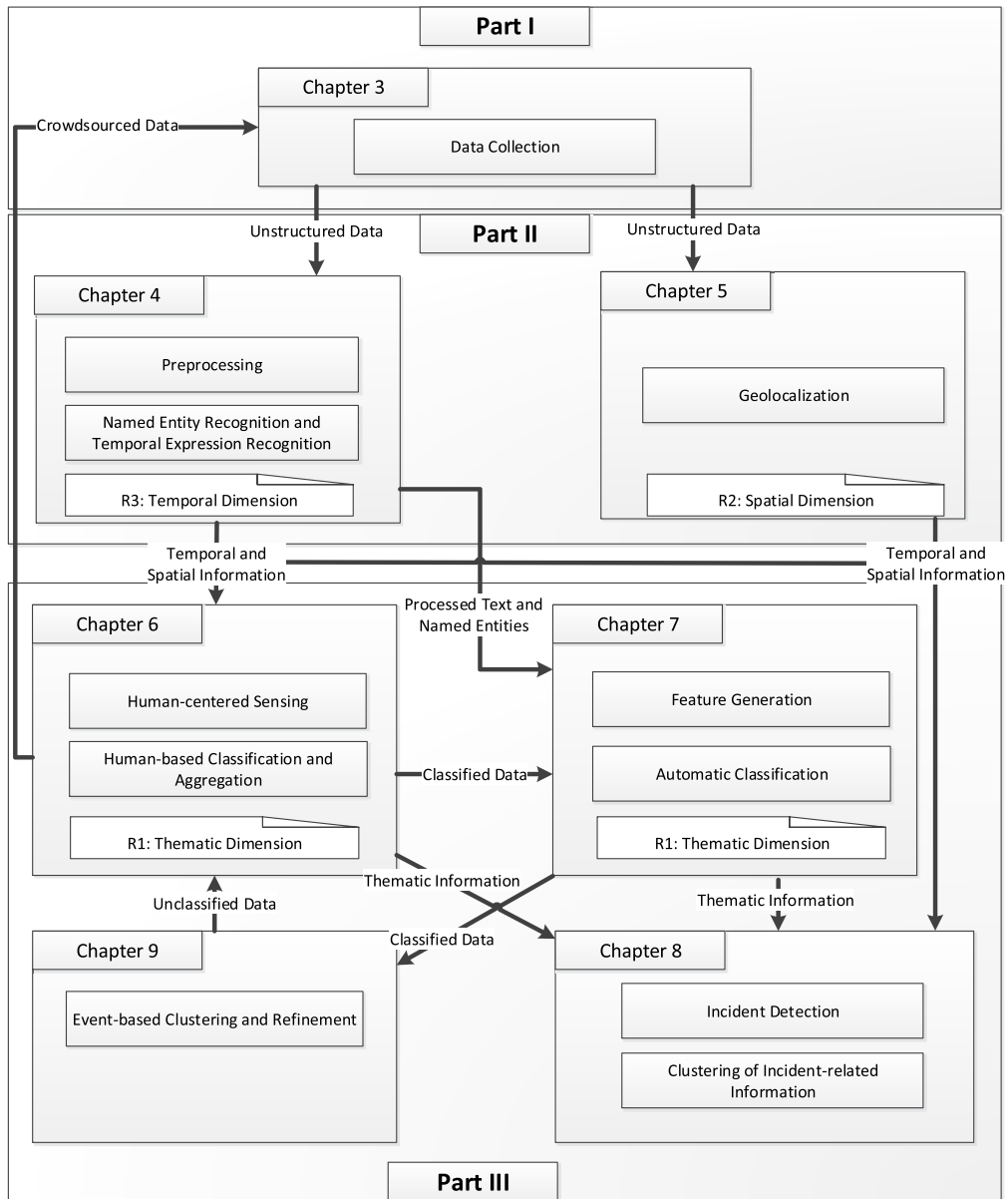


Figure 1.: Overview of the connections of the chapters in this dissertation.

Part I.

Initial Foundations for Detecting and Clustering Incident-Related Information in User-Generated Content



2 Design Considerations of a Framework for Detecting Incidents and Clustering Incident-related Information based on User-generated Content

In this chapter, we present the general architecture of a framework that allows detecting incidents and clustering incident-related information based on user-generated content. The framework consists of the necessary steps to help us answer the question *"How can user-generated content be made a usable and valuable source of information for situational awareness of decision makers?"*

In the first section, we define and characterize events and refine them for incidents as specific types of event (see Section 2.1). Based on this, we derive requirements for the framework. In the subsequent Section 2.2, we provide an overview of the framework.

2.1 Initial Definitions and Design Considerations

To deal with the research question, we first need to define an event and incident as a specific type of event. Based on these definitions, we infer requirements of a system for small-scale incident detection.

Event Definition

Up to now, there is no consensus on the definition of an event [153]. Thus, for this dissertation, we follow the most basic and general definitions of the term "event". In the first Topic Detection and Tracking (TDT) challenge, Allan et al. [13] defined an event as "some unique thing that happens at some point in time". This definition was later refined by Yang et al. [250] to "an event identifies something (non-trivial) happening in a certain place at a certain time". Both definitions show that an event is clearly characterized by spatial and temporal dimensions. Furthermore, events can be regarded as "instances of topics" [250]. In our case, a car crash that happened on March 12, 2014, on the highway A5 is an event, whereas "car crash" is a topic. Thus, an event is also characterized by a thematic dimension.

Thus, we finally associate an event with three basic properties:

- A topic (i.e., a thematic dimension)
- A location (i.e., a spatial dimension)
- A specific time period (i.e., a temporal dimension)

Based on these properties, we define an event as follows:

Definition I.1. An **event** is something that is happening in the real world at a certain place, at a certain time, and which can be described by a topic.

Incident Definition

Throughout this dissertation, we focus on *incidents* (or *accidents*) as a specific type of event that share the same three properties. Following the definition of an incident by the Federal Emergency Management Agency (FEMA) [28], we define an incident as follows:

Definition I.2. An **incident** is an unexpected event in the real world typically resulting in a damage or injury that happens at a certain place, at a certain time, and which can be described by a topic.

Furthermore, incidents can be further distinguished with respect to their impact. A disaster is a large-scale incident and has a different nature compared with a small-scale incident. According to the definitions presented by [28], large-scale incidents are extraordinary events resulting in an extensive involvement of organizations. This type of incident is likely to become a "trending event" as this type of event is rather uncommon and may affect many people. Thus, the frequency of published information about "trends" or "trending events" is substantially higher [122] compared with everyday events such as small-scale incidents. Small-scale incidents happen every day and everywhere and are typically of low public interest. For instance, an earthquake is likely to be a large-scale event because of the high absolute amount of information about this incident. Compared with this, for small-scale incidents, the amount of messages is not significantly higher as there are usually few reactions in social media because of a car accident.

Furthermore, both incident types differ with respect to spatio-temporal localization. Small-scale incidents have a small spatial and temporal extent, whereas the location of a large-scale incident is large or may not be well defined. For instance, a car crash might last for one hour and is limited to an intersection or a lane of a highway. In contrast, in 2012, the hurricane Sandy lasted for several days and affected many people at different locations.

Requirements of a System for Small-Scale Incident Detection

The goal of this dissertation is to differentiate incident-related user-generated content (i.e., every *information item* in user-generated content) from noise. According to the definition of an incident as a specific type of event, inferring the spatial, temporal, and thematic dimensions is necessary. As an information item is either related to a specific real-world incident or not, for detecting incidents and aggregating information related to the same incident, the three dimensions need to be inferred.

Thus, a system for small-scale incident detection needs to suffice the following three requirements:

R[1]: The system needs to derive the *thematic* dimension of an information item.

R[2]: The system needs to derive the *spatial* dimension of an information item.

R[3]: The system needs to derive the *temporal* dimension of an information item.

Based on the information derived for each information item, it can clearly be related to an incident. In this dissertation, we define an information item related to an incident as *incident report*. Based on the individual information of each incident report, information for the whole incident can be inferred. Furthermore, incident reports related to the same event can be clustered accordingly.

2.2 Architecture of a Framework for Small-Scale Incident Detection

In the following section, a framework for (1) detecting small-scale incidents based on user-generated content and for (2) clustering information related to the same incident is presented.

The framework relies on two approaches for analyzing a large amount of data: crowdsourcing (i.e., the engagement of humans for manual filtering of user-generated content) and machine learning for automatic extraction of useful information. The combination of both approaches is necessary as on the one hand, manual analysis of user-generated content is prone to errors. Furthermore, crowdsourcing might result in untrustworthy information [226]. Also, applying crowdsourcing in such time-critical situations as emergencies is not always applicable. On the other hand, machine learning algorithms need to be trained and validated. For this, annotated training data is needed. Also, one model trained on one city may not be applicable on data of a different city because of the nature of social media data. Thus, already trained models need to be refined to changing conditions. For the framework developed in this dissertation, we decided to combine both approaches to overcome the limitations of each individual one.

The six steps defined in the framework are summarized in the following:

1. **Collection and Filtering:** In the first and initial step, user-generated content is collected. Besides social media, where valuable information is directly provided, additional information from on-site bystanders can be collected using mobile applications. Furthermore, as the amount of data is large, a prefiltering of the information base can be applied. As a result of this step, a large amount of unstructured information is collected, which needs to be further processed.
2. **Automatic Preprocessing:** As the information obtained in the previous step is usually very short and contains noise, applying automatic processing steps such

as machine learning is difficult. Thus, in the second step of the framework, several automatic preprocessing steps are conducted. First, the unstructured information base is converted to a structured information base. Second, named entities and temporal expressions are identified to be used in subsequent steps. Based on the resulting information, the temporal and spatial dimensions for each information item is derived automatically (R[2] and R[3]).

3. **Human-Based Classification and Aggregation:** In this part of the framework, manual classification is applied to infer the thematic dimension of an incident (R[1]). Furthermore, manual classification is used to aggregate information related to the same incident. Also, additional information about an incident can be collected in this step. As a result, classified and aggregated incident reports are created.
4. **Machine-Based Classification:** As outlined before, crowdsourcing is limited when it comes to timely information retrieval on a large amount of user-generated content. Thus, based on the preclassified information, machine learning models are trained. These can be used to automatically infer the thematic dimension of an information item for a large amount of data (R[1]). As a result, incident reports are separated from information not related to incidents.
5. **Machine-Based Aggregation:** Finally, based on the spatial, temporal, and thematic information derived for each individual information item, incident reports can clearly be related to an incident. Based on this, new incidents can be detected. Also, reports related to the same incident are automatically clustered to provide a set of relevant information to a decision maker.
6. **Refinement:** As social media platforms are not static environments, the information base needs to be adapted to changing conditions. For instance, data from a different city or new information sources are used. Also, the dynamism of user-generated content might result in adapting the pipeline according to the current needs. Furthermore, decreasing classification quality may result in refining the automatic processing models used in the framework. In the refinement step, these adaptations are supported.
7. **Presentation and Usage:** As a result of this framework, a structured information base that enhances the situational awareness of a decision maker is created. The information can now be consumed and used for taking decisions. Furthermore, the resulting information can be fed again into the framework.

In Figure 2, the connection of each of the steps is shown.

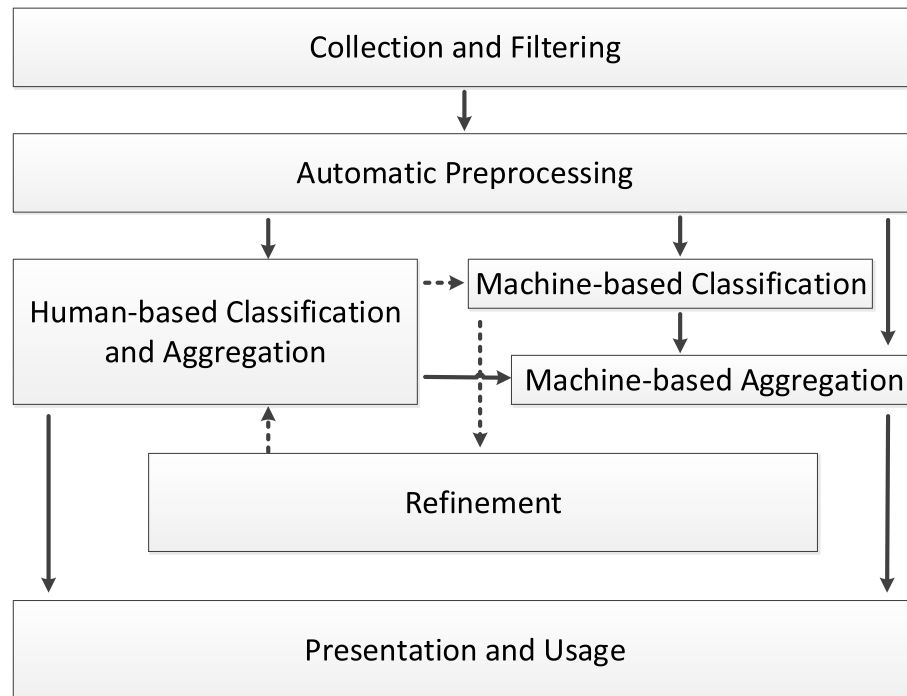


Figure 2.: Overview of the framework for small-scale incident detection

The first step, collection and filtering, creates the initial information base and filters all incoming data according to certain conditions such as the presence of certain keywords. All collected data is preprocessed, which in our case means that spatial and temporal information is extracted, which can directly be used in all subsequent steps for spatio-temporal filtering. Furthermore, unstructured data is converted into structured data to allow the inferencing of the thematic dimension.

Dependent of the input data, it is either used in the human-based classification and aggregation step, the machine-based classification step, or the machine-based aggregation step. In the first case, the input data is manually processed. As a result of this, classified and aggregated incident reports can directly be provided to a decision maker or can be used in the other steps of the framework. Furthermore, new information might be generated, which can then be used as new input for the framework. In the second case, the machine-based classification, each information item is automatically classified according to the thematic dimension. The output of this step can then be used in the machine-based aggregation step. This step relies on the spatial, temporal, and thematic dimensions inferred for each information item to detect incidents as well as to aggregate related items.

The refinement step is an iterative process that takes unlabeled data as input to select which information item is most valuable to annotate manually. Based on the manual annotation, the automatic classification in the machine-based classification step can be refined with new data. This process is continued until some stopping criteria such as a minimum level of classification accuracy is reached.

Finally, a large amount of previously unprocessed and unstructured data is now represented as a structured information base. This way, it can be used for decision making. Furthermore, data derived in the preceding steps can be used as new input for the framework. For instance, new information is collected in the human-based classification and aggregation step.

2.3 Conclusion

In this chapter, we defined and characterized an event and incident as a specific type of event. We showed that an incident is clearly defined by a spatial, temporal, and thematic dimension. Based on these dimensions, we identified requirements of a system that is able to identify incident-related information in user-generated content. We further presented the overall architecture of a framework sufficing the requirements and introduced the necessary processing steps.

In the following chapter, we present the first step of the pipeline, which is the collection and filtering of user-generated content.

3 Considerations of Collecting and Filtering User-Generated Content

In the last chapter, we presented the overall framework for incident detection. As a first step of the framework, user-generated content is collected and filtered (see Figure 3). As a result of the collection step, an initial information base is created, which can further be processed in the subsequent steps of the framework.

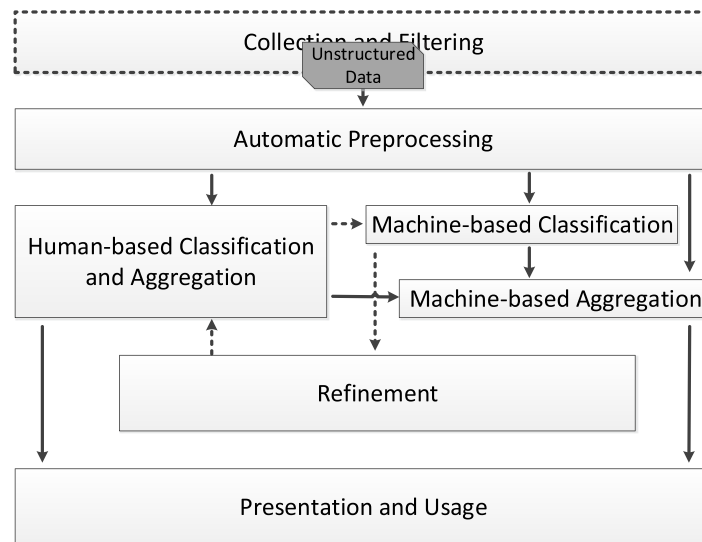


Figure 3.: Collection and filtering as the first step in the framework.

In the first section of this chapter, we present background on user-generated content and Twitter (see Section 3.1). In the second section, we present our data collection approach (see Section 3.2). The chapter is summarized in Section 3.3.

3.1 Background on User-Generated Content and Twitter

User-generated content is defined by [118] as "various forms of media content that are publicly available and created by end-users." Social media was built upon the principles of the Web 2.0 and allows the sharing and creation of user-generated content. Different types of social media [91] can be differentiated: social networking sites such as Facebook¹ and LinkedIn² allow connecting with friends, whereas social

¹ <https://www.facebook.com/> [Accessed: 15.01.2014]

² <https://www.linkedin.com/> [Accessed: 15.01.2014]

media data in the form of videos, audio files, or photos is shared on YouTube³ or Flickr⁴. Textual content is mostly shared in blogs such as Engadget⁵ or with limited content on microblogging sites such as Twitter⁶ or Tumblr⁷.

In this dissertation, we focus on textual content as there has been a rapid growth of text data in social media [6]. Furthermore, we focus on Twitter as one very prominent platform on which information is shared every day and by a variety of people. In 2013, Twitter had about 240 million active users [232], who shared more than 500 million messages per day [205]. This huge amount of data gives a wide base of information for a variety of topics.

On Twitter, users can post short messages called *tweets* of up to 140 characters in length. These *microposts* are either sent from mobile devices, from third-party applications, or from web applications. For each user, a stream of microposts is displayed as a *microblog*, which is the reason why Twitter is a microblogging platform. Twitter is also a social network as users are able to follow each other's microblogs. Furthermore, users can forward or *retweet* each other's messages.

While communicating, people use a variety of Twitter-specific symbols [125]. Place-names or user names are referenced using the "@" symbol. Also, Twitter allows to use the hashtag "#" symbol to specify a number of keywords or a topic of a tweet. For instance, "#swineflu" was introduced for the trending news event. However, there is no common convention on how to name these topics [44]. Furthermore, hashtags are not necessarily unique and are highly dependent on how they are used in the whole social network [188].

As outlined in the introduction, social media data such as tweets is inherently noisy and unstructured. In Listing 3.1, an example tweet that illustrates the unstructuredness of textual information in social media is shown.

Listing 3.1: Example tweet showing the unstructuredness of textual information.

```
RT: @People - 0noe friday afternoon in heavy traffic, car crash on I
-90, right lane closed
```

First, Twitter-specific annotations such as @-mentions and retweet annotations are used. Second, abbreviations such as "0noe" are present. Third, very short sentences are written due to the restricted length of a tweet. However, the information density is high as, for example, the position, and the type of incident is mentioned.

³ <https://www.youtube.com/> [Accessed: 15.01.2014]

⁴ <https://www.flickr.com/> [Accessed: 15.01.2014]

⁵ <http://www.engadget.com/> [Accessed: 15.01.2014]

⁶ <https://twitter.com/> [Accessed: 15.01.2014]

⁷ <https://www.tumblr.com/> [Accessed: 15.01.2014]

3.2 Data Collection and Filtering of User-Generated Content

In the following, we describe how user-generated content is collected. For this, we introduce two Twitter application programming interfaces (APIs) for gathering tweets.

Twitter provides two major APIs for collecting tweets. First, the Streaming API⁸ allows the crawling of real-time Twitter data. This API gives a 1% real-time stream of all tweets created worldwide. Second, the Search API⁹ can be used to get tweets related to certain keywords or a location. It allows specifying a search query containing multiple keywords and GPS coordinates as well as a radius. Using this API, it is possible to collect a stream of tweets for a single city. The Search API provides not only explicitly geotagged tweets but also tweets that have been geocoded by Twitter (e.g., using the user profile). However, the results provided by this API are not complete sets of all tweets, but they are prefiltered by Twitter¹⁰.

In the following, we give an overview of data sets created with both APIs. Furthermore, as the amount of data is not applicable for manual analysis, it needs further filtering. For this, we present our approach for keyword filtering.

3.2.1 Data Sets

We collected four large tweet data sets, which are the base for our evaluations in this dissertation. In the following, we present the initial data sets that were collected. In each chapter, we provide a detailed description of how these data sets were adapted to suffice the specific needs of the respective evaluation. Also, in the following chapters, additional data sets that are only used in the respective chapters are introduced.

From September 2011 to February 2012, we crawled around 80 million tweets from the so-called *Spritzer* stream using the Twitter Streaming API. The data set that we collected was used as a worldwide data set for developing the approach for inferring the spatial dimension of an information item (see Chapter 5). The resulting data set is as follows:

- **SET_GEO:** 80 million worldwide tweets collected from September 2011 to February 2012.

Furthermore, we collected several city-specific data sets using the Search API. These data sets were collected in a 15 km radius around the city centers of Seattle, Wash-

⁸ <https://dev.twitter.com/docs/streaming-apis/streams/public> [Accessed: 15.01.2014]

⁹ <https://dev.twitter.com/docs/api/1.1/get/search/tweets> [Accessed: 15.01.2014]

¹⁰ <https://dev.twitter.com/docs/faq> [Accessed: 15.01.2014]

ington, Memphis, Tennessee, New York City, New York, Chicago, Illinois, and San Francisco, California. We focused on these cities as they have a huge regional distance, which allows us to evaluate our approaches with respect to geographical variations. Furthermore, 15 km was chosen because with this radius, we are able to get data even for larger cities.

Though we know about the limitations of the Search API, using the Search API has been shown as the most appropriate means for analyzing tweets in prior work [236]. As the Search API provides not only explicitly geotagged tweets but also tweets that have been geocoded by Twitter, we assume that we retrieved a relevant sample for our experiments, although it is not a complete set for the cities.

We collected all tweets that were available using the Search API during certain time periods for the cities. This gave us three initial sets of tweets, which are used in several chapters of this dissertation:

- **SET_CITY_1:** 6M tweets collected from November 19, 2012 to December 19, 2012 for Memphis, Tennessee and Seattle, Washington.
- **SET_CITY_2:** 1.5M tweets collected from February 1, 2013 to February 7, 2013 for Memphis, Tennessee and Seattle, Washington.
- **SET_CITY_3:** 2.5M tweets collected from January 1, 2014 to March 11, 2014 for New York City, New York; Chicago, Illinois; and San Francisco, California.

3.2.2 Incident Types and Keyword Filtering

As we needed to restrict our research scope, we decided to focus on a subset of incident types. In the following, we present which types were chosen to be used in this dissertation. We decided to focus on three specific incident types because we identified them as the most common incident types in the city of Seattle, Washington using the Seattle Real Time Fire Calls data set¹¹. The Fire Calls data set is a source of frequently updated official incident information and provides high-quality data about incidents.

Thus, in this dissertation, we focus on very common and distinct incident types and one neutral event type:

- Car incident
- Fire incident
- Shooting incident
- Not incident related or other type of incident

¹¹ <http://seattle.data.gov> [Accessed: 01.03.2014]

Whenever a manual analysis of data sets was applied, we needed to reduce the initial data sets to ensure high-quality ground truth data for our experiments. Thus, our approach used for filtering is to identify and extract tweets mentioning incident-related keywords. These keywords are derived from the incident types. Though keyword filtering significantly reduces the overall amount of data, it helps to build an initial set for developing and optimizing algorithms. The keyword filtering is applied on the complete set of tweets; thus, the keywords are identified after collecting all available tweets with the respective API. Compared with other approaches that completely rely on filtering using hashtags, we take the whole message into account for identifying incident-related keywords.

To build a set of incident-related keywords, we retrieved all incident types used in the "Seattle Real Time Fire 911 Calls" data set. Based on these incident types, we manually identified those that matched our three incident types. This gave us a set of 11 incident types that stem from the real-world information retrieved from Seattle. For each of these incident types, we defined one general keyword set with keywords that are used in the name of the incident type. For instance, the keywords "vehicle", "accident", and "vehicle accident" were assigned as general keywords for the incident type "Motor Vehicle Accident".

The general keywords were then further enhanced with a set of additional keywords. These specific keywords were identified by extending the general keywords with the direct hyponyms extracted from WordNet¹². For instance, the keyword "accident" was extended with the related words "collision", "crash", "wreck", "shipwreck", "injury", "accidental injury", "fatal accident", and "casualty".

An overview of the incident types and the overall number of extracted keywords can be found in Table 1. Based on these 257 incident-related keywords, we are able to prefilter tweets to a set of information that is likely related to incidents.

¹² <http://wordnet.princeton.edu> [Accessed: 15.01.2014]

Table 1.: Real-world incident types and number of extracted keywords.

Incident Type	Fire Incident	Shooting Incident
Real-world incident types	Fire In Building	Assault w/Weapon
	Fire In Single Family Residence	Assault w/Weapons-Aid
	Automatic Fire Alarm Residence	
	Auto Fire Alarm	
# of keywords	148	36
Car Incident		
Real-world incident types	Motor Vehicle Accident	
	Motor Vehicle Accident Freeway	
	Aid Response Freeway	
	Car Fire	
	Car Fire Freeway	
# of keywords	73	

3.3 Conclusion

In this chapter, we presented the first step of our framework, which is the collection of user-generated content. We gave an overview of user-generated content and Twitter, as the platform that our analyses are based on. We further presented our data collection approach and introduced data sets that are used throughout this dissertation. Also, we introduced three specific incident types as well as a keyword-filtering approach to reduce incoming user-generated content to a set of manageable information. As a result of the collection step, an initial information base is created, which can further be processed in the subsequent steps of the framework.

As the data collected in this step is unstructured, the following part deals with the problem of how to preprocess user-generated content.

Part II.

Automatic Preprocessing and Geolocalization of User-Generated Content



The last part provided an overview of the framework for small-scale incident detection. Furthermore, we introduced the first step of the framework for collecting user-generated content. As the collected texts shared in social media are mostly unstructured, further processing is needed. The following part presents preliminaries needed for inferring the thematic dimension of an event in further processing steps (see Figure 4). Also, we show how we infer the temporal and spatial dimensions of a tweet.

As one of our requirements is to infer the thematic dimension of an event (R[1]), unstructured texts need to be preprocessed so they can be used as structured data for feature generation. The necessary preprocessing steps are presented in Chapter 4. In the same chapter, we show how named entities and temporal mentions are identified and extracted so that they can be used as additional features when applying data mining. For this, in Section 4.2, we describe how we adapted existing approaches for named entity and temporal expression recognition. We also present how we make use of the temporal expressions to infer the temporal dimension of a tweet (R[3]).

In the second chapter of this part (see Chapter 5), we deal with the problem of how we infer the spatial dimensions of a tweet (R[3]). In that chapter, we present the first major contribution of this dissertation, which is a novel approach for the geolocalization of user-generated content.

The contributions presented in this part are the following:

- We present which preprocessing steps are conducted to create structured text that can be used for feature generation.
- We present a set of adaptations applied to standard techniques that allow us to extract named entities and temporal expressions from unstructured text. Furthermore, these approaches are evaluated on data sets of Twitter messages.
- We propose an adaptation of a standard approach for extracting temporal expressions and show how we make use of the results for inferring the temporal dimension of an event mentioned in a tweet.
- We propose a novel approach for the geolocalization of tweets, suitable for inferring the home location of a Twitter user, the point of origin where a tweet was sent, as well as for inferring the location focus of a tweet message. In an evaluation, we show that the approach is able to locate 92% of all tweets with a median accuracy of below 30 km. Furthermore, it predicts the user's residence with a median accuracy of below 5.1 km. Finally, the same approach is able to estimate the focus of incident-related tweets within a median accuracy of below 250 m.

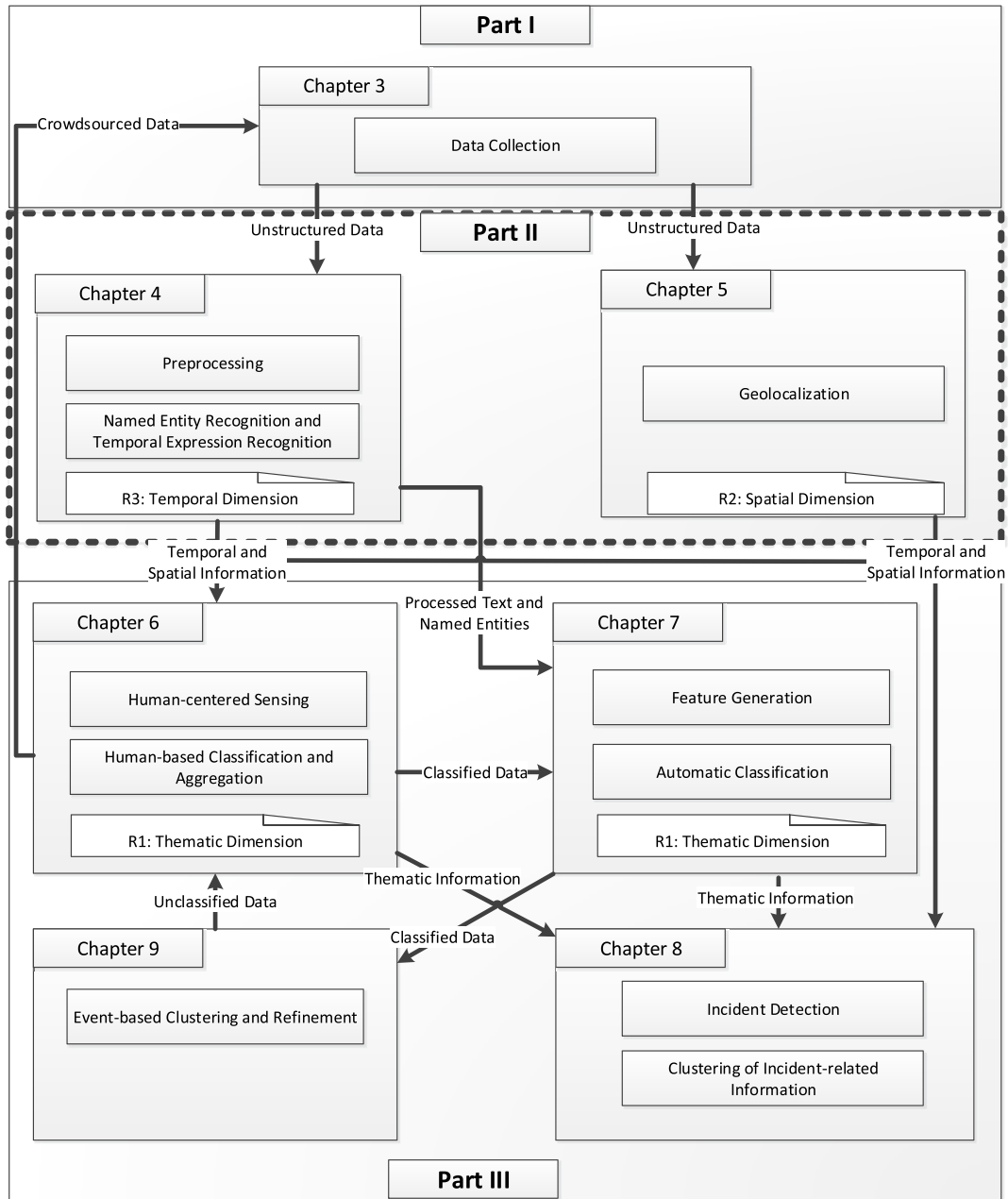


Figure 4.: Overview of the connections of Part II to the chapters in this dissertation.

4 Preprocessing of Unstructured Text

In this chapter, we provide the preliminaries needed for inferring the thematic, spatial, and temporal dimensions of a tweet. For this, we present several approaches that are conducted as preprocessing steps on the unstructured Twitter message. As a result of this preprocessing step, structured text is created. Furthermore, named entities and temporal information are identified (see Figure 5).

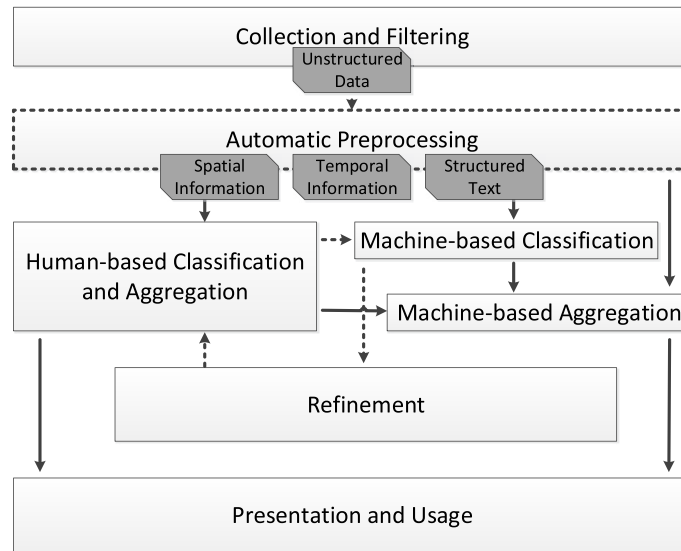


Figure 5.: Automatic preprocessing as a second step of the overall framework.

To detect incident-related information in user-generated content, we aim to use a machine learning model. Training this model requires documents in a manageable representation (i.e., a feature vector). For creating this feature vector, documents need to be converted into a representation that enables the creation of this feature vector. This conversion is much more difficult compared with using structured documents as texts in social media are unstructured. Thus, related to the requirement of inferring the thematic dimension [R1], we need to answer the following question:

- *Which preprocessing steps are needed to convert unstructured to structured text for feature generation to allow inferring the thematic dimension?*

To this aim, we first present which Natural Language Processing steps are applied on social media data (see Section 4.1).

Furthermore, named entities present in social media texts provide additional information about a geolocation and the point in time when an event occurred, which are important for incident detection ([R2] and [R3]). Also, named entities can be used to derive additional features for the text mining process in addition to the plain word-based approach (see Chapter 7). For this, temporal and spatial information need to be identified in the unstructured data, which is covered by the following two questions:

- *How is spatial information identified in social media texts?*
- *How are temporal expressions identified in social media texts?*

Related to these questions, we show how we detect named entities and temporal expressions present in textual documents in Section 4.2. Furthermore, in the same section, we present how we infer the temporal dimension of a tweet based on the extracted temporal expressions.

In this chapter, we present several adaptations of standard techniques:

- We present which preprocessing steps are conducted to create structured text, which can be used for feature generation.
- We present a set of adaptations applied to standard techniques that allow us to extract named entities and temporal expressions from unstructured text. Furthermore, these approaches are evaluated on data sets of Twitter messages.
- We propose an adaptation of a standard approach for extracting temporal expressions and show how we make use of the results for inferring the temporal dimension of tweets.

Though the preprocessing steps are explained using tweets as examples, they are likewise applicable to different unstructured texts that stem from social media.

In the first section of this chapter, we present which preprocessing steps are conducted on the unstructured text to convert it into a structured representation (see Section 4.1). In the second section, we show how we extract named entities and temporal expressions (see Section 4.2). Furthermore, we show how we infer the date of an event mentioned in a tweet. The results are summarized in Section 4.3.

Parts of this chapter appeared in [201].

4.1 Natural Language Preprocessing

In this section, we discuss which Natural Language Processing steps must be applied on social media data to structure unstructured text to finally enable the creation of a feature vector. In Listing 4.1, an example of an unstructured text in a Twitter message is shown.

Listing 4.1: Example of unstructured text sent in a tweet.

RT: @People - @noe friday afternoon in heavy traffic, car crash on I
-90, right lane closed

Manning et al. [145] presented four preprocessing steps that are conducted on regular text documents to allow feature generation:

1. Collection of documents
2. Tokenization
3. Normalization
4. Feature vector creation

Whereas the first step was conducted in Chapter 3 and the last step is part of Chapter 7, tokenization and normalization are part of our preprocessing steps.

After collecting an initial set of tweets, we tokenize them. Thus, the text is divided into discrete words (*tokens*) based on different delimiters such as white spaces. Every token is then analyzed, and nonalphanumeric characters are removed or replaced. Next, the resulting text is normalized. The goal of the normalization is to derive a common base form for a word. For instance, the words "am" and "are" are derived from "be". For this, we apply lemmatization to reduce each word to a common base form (i.e., a lemma). We decided to use lemmatization instead of stemming as stemming is a simple heuristic process which may result in inappropriate stems [145].

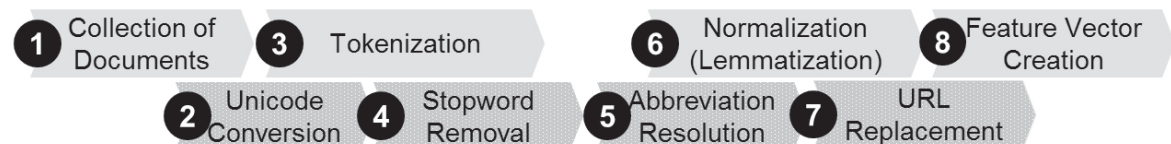


Figure 6.: Extended preprocessing pipeline for social media texts based on the steps proposed by Manning et al. [145].

As we have to deal with social media texts, we enhance the common preprocessing steps with additional ones (see Figure 6).

- As a first step and before further processing is applied, the text is converted to Unicode. This step is needed as some tweets contain non-Unicode characters.
- As a next step, stopwords are removed. Stopwords are very frequent tokens such as "the". Due to their high frequency, these words have limited influence when it comes to applying machine learning on tweets.
- As abbreviations and slang are commonly used in social media data [183], we replace them with the resolved abbreviations. For this, we use a dictionary

provided by the Internet Slang Dictionary & Translator¹³. For instance, in the example tweet, "Onoe" is replaced with "oh no".

- URLs are widely used in social media for linking to external websites. However, this results in a variety of different URLs that would be added as single features to the feature vector. Thus, we replace them with one, common token "URL".

Whereas the value of Unicode conversion and stopwords removal is obvious, in Chapter 7, we evaluate the advantage of the abbreviation resolution and URL replacement steps.

In Listing 4.2, the preprocessed example tweet after applying the preprocessing steps of Figure 6 is shown. Based on this preprocessed text, we are able to generate a feature vector for text mining, which is applied in the following steps of the framework.

Listing 4.2: Preprocessed example tweet.

```
friday afternoon traffic car crash I-90 lane close
```

4.2 Named Entity and Temporal Expression Recognition on Unstructured Text

In the last section, we described an extended pipeline to generate structured from unstructured text. Additionally, we perform several problem-dependent preprocessing steps. Based on the unprocessed text, we identify named entities and expressions to use the extracted information (1) as additional features for text mining and (2) for incident detection. The results of these steps serve as input for the rest of the processing pipeline. First, named entities can be used for feature generation to finally detect the type of incident ([R1]). Second, we infer additional information about the spatial and temporal dimensions of an incident ([R2] and [R3]).

In the following, we present adaptations of standard approaches for identifying and classifying named entities and temporal expressions in tweets. These steps are performed on the unprocessed tweet in parallel to the preprocessing steps shown in the previous section.

1. We show how we use Linked Open Data (LOD) as a source of interlinked information about various types of entities (see Section 4.2.2). With the presented steps, we are able to generate new features for data mining.
2. Named Entity Recognition (NER) is applied to extract location mentions (see Section 4.2.3). Location mentions are used to generate new features. Also, they are valuable for the geolocalization of user-generated content.

¹³ <http://www.noslang.com/> [Accessed: 19.01.2013]

-
-
3. We present a framework for temporal expression extraction (see Section 4.2.4). Temporal expressions are also used to generate new features for text mining. Also, they can be used to infer the point in time of the event mentioned in a tweet.

These preprocessing steps are finally evaluated in Section 4.2.5.

4.2.1 Definition of Named Entities and Temporal Expressions

Named Entity Recognition (NER) is the task of identifying and extracting expressions (or mentions) in text [158]. Furthermore, these mentions are classified according to predefined types (e.g., organization names, location names, or person names). Named entities were initially defined as "unique identifiers of entities" [46]. However, up to now, there is no common agreement on the definition of a named entity in the research community [147]. Thus, we use the following definition throughout this dissertation, which extends the definition of Petasis et al. [172]:

Definition II.1. An **entity** is a thing that can be uniquely identified by its properties (e.g., United Kingdom, Seattle, my university). A **named entity** is an entity that has been assigned a name (Technische Universität Darmstadt). Thus, the mention of a named entity in a text is defined as **named entity mention**.¹⁴

We further distinguish named entities of type location as a specific type of named entities:

Definition II.2. A **proper location mention** (also called **toponym**) is defined as the named entity mention of a location.

Typically, a location mention is a proper name (represented by a noun or noun phrase) was given to a location. In contrast, common location mentions are defined as follows:

Definition II.3. **Common location mentions** are location mentions for which no indication of the name is given in a text.

For example, "Seattle" is a proper location mention, whereas "my university" in "I go to my university." gives no indication of the name of the university. Thus, the latter one is a (common) location mention.

Temporal expressions are another important part of texts. By definition, they are not named entities; thus, beside NER we apply Temporal Expression Recognition and Normalization (TERN) [8]. TERN copes with detecting and interpreting temporal

¹⁴ In the following, we use quotes to distinguish words from entities, which are written in italics.

expressions to allow further processing. For example, in our case, we use temporal expressions to determine the point in time of an event. According to the definition of Ahn, Van Rantwijk, and De Rijke [8], we define temporal expressions as follows:

Definition II.4. *Temporal expressions* are tokens or phrases in text that serve to identify a point in time.

According to this definition, the phrases "yesterday", "last Monday", "05.03.2013", or "2 hours" are all temporal expressions.

4.2.2 Named Entity Recognition and Replacement Using Linked Open Data

We use Linked Open Data (LOD) as a source of interlinked information about entities. Tim Berners-Lee introduced Linked Data with four design principles [26]:

1. URIs have to be used as names for things.
2. HTTP URIs have to be provided so that people can look up the names.
3. Useful information needs to be provided when a URI lookup is performed, using the standards (e.g., the Resource Description Framework (RDF) or SPARQL).
4. Links to other URIs need to be provided as additional resources.

In our case, each (named) entity is denoted by a unique uniform resource identifier (URI) of the form `http://dbpedia.org/resource/Name`. For instance, in DBpedia, the named entity *New York City* is denoted by the URI `http://dbpedia.org/page/New_York_City`. Also, every entity is connected with other entities, which is important when it comes to using external knowledge about entities.

To represent named entities from social media texts as Linked Data, we stick to the first and second principles and identify a HTTP URI for each named entity. Later on, we show how we use links to other entities (see Chapter 7) according to the fourth principle.

In this dissertation, we use DBpedia as a source of LOD. In the DBpedia [27] project, information from Wikipedia¹⁵, YAGO¹⁶, and other sources was extracted and provided as interlinked information. Furthermore, we use two relations that are present in DBpedia. First, similar entities are grouped into classes expressed by a type relationship to the URI of a class or the resource describing the class respectively. For example, *New York City* is an entity of type *City*, which itself has its own URI, `http://live.dbpedia.org/ontology/City`, in LOD. Another important relation is `http://purl.org/dc/terms/subject`, which relates to one or many resources describing the topic of an entity.

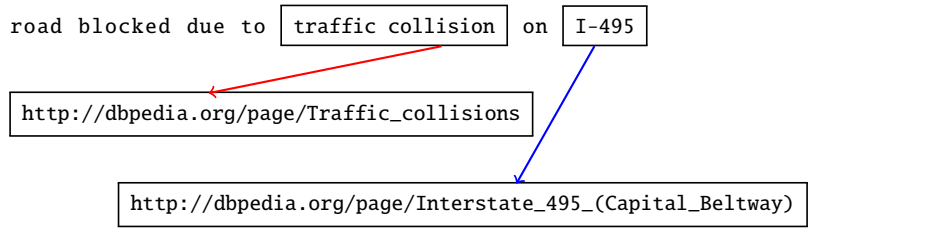
¹⁵ <http://de.wikipedia.org/> [Accessed: 19.02.2013]

¹⁶ <http://www.mpi-inf.mpg.de/yago-naga/yago/> [Accessed: 19.02.2013]

To identify named entities and map them to URIs in DBpedia, several APIs have been evaluated in related work (see, for examples, [185, 83, 220]). *DBpedia Spotlight* [155] has shown good performance for NER [185]. Furthermore, Spotlight is available without cost and has an unrestricted number of API calls, which is why we decided to use this framework.

Spotlight provides an entity recognition algorithm, which is able to disambiguate named entity mentions based on information provided in the rest of the text. We use the Spotlight API without any adaptations. The results provided by Spotlight are used in Chapter 7 as it allows for a semantic abstraction from the concrete instances a tweet talks about, which is valuable for text classification. In Listing 4.3, HTTP URIs extracted for an example tweet are shown, which is the result of extracting named entities in the preprocessing step using Spotlight:

Listing 4.3: Extracted DBpedia properties for a tweet.



In Chapter 5, we make use of this extraction approach for the geolocalization of user-generated content ([R2]). Furthermore, in Chapter 7, we show how these entities are used for inferring the thematic dimension of a tweet ([R1]).

4.2.3 Location Mention Extraction and Replacement

In the following, we present how we extract location mentions, which are valuable for the geolocalization of user-generated content (see Section 5) as well as for generating new features for data mining. We found that especially common location mentions are used rather frequently in incident-related tweets. For instance, during our analyses of incident-related tweets, we often found geospatial entities such as "highway" or "school".

The first tweet shown in Listing 4.4 gives an example of how location mentions are used. In the example tweet, the proper location mention "I-90" is used. With the following approach, we aim to recognize these location mentions. This extraction includes different named entities such as streets, highways, landmarks, or blocks.


For location mention extraction and replacement, we use the Stanford Named Entity Recognizer [74] to identify location mentions in tweets. We decided to use this

framework as it showed good performance in related approaches [5]. The Stanford NER is based on Conditional Random Fields (CRFs)¹⁷.

We retrained the recognizer on tweets to use the adapted model to annotate location mentions in tweets as shown in Listing 4.4. Based on this, mentions are detected and replaced with a general annotation "ProperLOC". We also detect common location mentions such as "home", "office", or "school" and replace them with a general annotation "CommonLOC".

Listing 4.4: Example tweet with location mention and the same tweet with the replaced location mention.

```
RT: @People - Onoe friday afternoon in heavy traffic, car crash on
    I-90 , right lane closed
```



```
RT: @People - Onoe friday afternoon in heavy traffic, car crash on
    ProperLOC , right lane closed
```

In Chapter 5, we make use of this extraction approach for the geolocalization of user-generated content ([R2]). Furthermore, in Chapter 7, we show how these entities are used for inferring the type of incident ([R1]).

4.2.4 Temporal Expression Recognition and Normalization on Unstructured Text

In the preprocessing step of the framework, we also extract temporal expressions from tweets, which enables us (1) to use this information as features for text mining ([R1]) and (2) to infer the point in time of an event mentioned in a tweet. For example, the tweet shown in Listing 4.5 contains the temporal expression "friday afternoon" referring to the point in time when an accident occurred.

For identifying temporal expressions in texts, several frameworks have been proposed [221, 24, 234, 138]. However, none of the existing approaches have been applied to tweets. For identifying temporal expressions in tweets, we decided to adapt the HeidelTime [221] framework. The framework has been chosen because the authors showed good performance on various data sets [235]. Furthermore, the framework is easily extensible. The HeidelTime framework uses regular expressions to detect temporal expressions in texts. As the system was developed for large text documents with formal English language, it was not optimized to detect temporal expressions in tweets.


¹⁷ See [126] for a comprehensive introduction

We made the following adaptations to provide good results on short and unstructured texts. First, abbreviations and slang are resolved. This is necessary as, for instance, Ritter et al. [183] showed that more than 50 variances of the word "tomorrow" are used in social media texts such as "2marrow", "2maro", "tomrw". As mentioned in Section 4.1, we use a dictionary for resolving commonly used abbreviations and slang.

Second, we extended the standard HeidelTime tagging functionality to annotate temporal expressions such as dates and times¹⁸ with two annotations: "DATE" and "TIME". As a result, the temporal expression in the example tweet is replaced with our annotation (see Listing 4.5).

Listing 4.5: Replaced temporal expression in example tweet.

```
RT: @People - @noe friday afternoon in heavy traffic, car crash on I
-90, right lane closed
```



```
RT: @People - @noe @DATE in heavy traffic, car crash on I-90, right
lane closed
```

Third, the annotated temporal expressions are used to provide an estimation of the point in time when an event mentioned in a tweet occurred. This is important as using the creation date of a tweet is not always correct as people also report on incidents that occurred in the past. For estimating this point in time, we use the creation date of a tweet as the base for our estimations. Using the extension, all temporal expressions are extracted and combined with the creation date to calculate the date when the event could have occurred. The result is finally returned in a machine-readable format.

For instance, given the tweet created on Friday 15, 2013, 14:33 (see Listing 4.6), we can use our mechanism to estimate February 14, 2013, as a likely point in time when the accident occurred as the tweet is referring to "yesterday".

Listing 4.6: Example tweet for determining likely point in time of an event.

```
Yesterday, two people died in a car accident. [created at: 15.02.2013,
14:33]
```

In Chapter 7, we show how these temporal expressions are used for inferring the type of incident ([R1]). Furthermore, based on this approach, we are able to detect the point in time when an incident occurred [R3].

¹⁸ Durations are not used as they are not valuable for detecting the time when an incident occurred.

4.2.5 Evaluation of Named Entity and Temporal Expression Recognition

In the following section, we present several studies showing the performance of our approaches for named entity and temporal expression recognition. The goal of these studies is to show that named entity and temporal expression recognition is possible with high performance on social media texts. We also evaluate the performance of detecting the point in time when an event occurred, which is important for creating a system that suffices R[3].

4.2.5.1 Data Sets and Metrics

In the following, we give an overview of the data sets and metrics used for our evaluations.

Data Sets

As there are no public incident-related tweet data sets available for our studies, we created data sets based on SET_CITY_1 and SET_CITY_2, which were collected using the Twitter Search API (see Section 3.2.1). We applied the incident keyword filtering presented in Section 3.2.2 on both data sets to identify tweets that contain incident keywords. From the resulting set, we randomly selected 2,000 tweets containing at least one incident-related keyword.

- **SET_SPOT:** For evaluating the recognition rate of named entities on tweets we used the complete data set consisting of 2,000 tweets. No further labeling of the data set was conducted.
- **SET_LOC:** For evaluating the recognition rate of location mentions, two researchers of our department annotated the 2,000 tweets with location mentions. Each location mention was assigned one of three labels: ProperLOC, CommonLOC, and OTHER.
- **SET_TEMP:** The evaluation of the recognition rate of temporal expressions was conducted on 500 randomly selected tweets of the 2,000 tweets. Each temporal expression in each tweet was labeled by one researcher of our department.
- **SET_EVENT:** For evaluating the recognition rate of a point in time of an event, we used 100 randomly selected tweets of the 2,000 tweets. Two researchers identified a likely point in time a tweet refers to and assigned this as ground truth data. If no explicit temporal information was contained in the message, the creation date of the tweet was used as ground truth.

Metrics

In the evaluations presented in this section, we provide metrics commonly used in information retrieval [245]:

$$\text{Accuracy (ACC)} = \frac{\text{Number of correctly classified named entities}}{\text{Total number of named entities}} \quad (1)$$

$$\text{Precision (P)} = \frac{\text{Correctly classified named entities}}{\text{Total predicted as named entities of a certain type}} \quad (2)$$

$$\text{Recall (R)} = \frac{\text{Correctly classified named entities}}{\text{Total number of named entities of a certain type}} \quad (3)$$

$$\text{(balanced) F-measure (F)} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

Furthermore, to get an understanding of how well the approaches perform on tweets, we provide the *Named Entity Recognition Rate* (i.e., the percentage of all tweets for which named entities can be extracted).

4.2.5.2 Results

In the following, we present the results of four studies. In the first three studies, we analyzed how well our approaches perform for recognizing named entities and temporal expressions. The last study was conducted to evaluate the performance of detecting a point in time when an event mentioned in a Twitter message occurred.

Study 1: Named Entity Recognition Rate Using Spotlight

To get an understanding of how well Spotlight performs on tweets, we evaluated the Named Entity Recognition Rate. We are only interested in the recognition rate as it gives an indicator if Spotlight is a suitable means for mapping named entities to URIs. We do not provide an in-depth evaluation of precision and recall as there is no ground truth data set available to determine if entities are correctly linked. For the study, we used data set SET_SPOT.

For determining the rate, we used the standard parameter settings for Spotlight (0.2 Confidence, 20 Support, 0.2 Contextual Score). In Table 2, an overview of the number of tweets for which named entities can be detected is given. The results show that we are able to detect named entities in about 57% of our data set; thus,

many potentially valuable named entities can be extracted. Furthermore, 1,871 mappings to URIs could be established. Hence, we assume that Spotlight is a suitable means for interlinking named entities in tweets with DBpedia.

Table 2.: Named entity recognition rate for tweets using Spotlight.

	Recognition Rate
SET_SPOT	57.10%

Study 2: Location Mention Recognition Rate

In a second study, we evaluated the quality of the location mention extraction approach. For this, we retrained the Stanford NER model on 1,200 tweets of SET_LOC and tested it on the remaining 800 tweets. Based on these data sets, we evaluated the models for detecting proper as well as common location mentions. With the evaluation results, we try to provide a rough understanding of how the accuracy of the model behaves when it is trained on mostly unstructured tweets.

The evaluation results are shown in Table 3. The results indicate that the retrained Stanford NER model shows high precision and recall (P=94.20% and R=91.29%) for recognizing location mentions in tweets. Thus, it is suitable for recognizing location mentions in tweets.

Table 3.: Evaluation results of location mention extraction approach.

Approach	Accuracy	Precision	Recall	F-measure
Stanford NER on Tweets	95.57%	94.20%	91.29%	93.71%

Study 3: Temporal Expression Recognition Rate

In a third study, we evaluated the recognition rate of temporal expressions in tweets. For the evaluation, we used the 500 tweets of SET_TEMP. The adapted HeidelTime framework was applied to identify these temporal expressions. The results of the evaluation are shown in Table 4.

Table 4.: Evaluation results for temporal expression extraction.

	Accuracy	Precision	Recall	F-measure
Adapted HeidelTime	95.6%	92.5%	91.3%	91.9%

The results indicate that the extraction of temporal expressions is possible with high precision and recall. However, several expressions are not extracted, such as "8:15". The example could be an expression of the score of a sports game, but also a temporal expression. This problem shows that disambiguation of temporal expressions

is not easily achieved. For resolving this problem, the rules of HeidelbergTime could be adapted in future work to detect these special cases. However, the current approach performs sufficiently to be used throughout this dissertation.

Study 4: Inferring the Point in Time of an Event

As it is essential to detect the point in time of an event [R3], we conducted a fourth study on this aspect. For temporal inference, we also used the adapted HeidelbergTime framework and applied it on SET_EVENT. The inferred point in time was then compared with the manually extracted point in time for each of the tweets. We only provide the accuracy as we are interested in the number of the correctly classified event dates of all ground truth dates.

The evaluation shows that we are able to detect the correct time and date of the event for 89 tweets. This gives us an accuracy of 89%. Though the approach shows good performance, it fails to extract the correct time for some tweets that contain multiple temporal expressions that refer to different time periods. For instance, the following tweet (see Listing 4.7) contains "Last Wednesday" as well as "10 years since" as temporal expressions. Both types of expressions are detected and extracted with our approach. However, the final date is calculated as February 20, 2013, based on the creation date and "Last Wednesday". The correct date would be the event that happened 10 years ago; thus, the true date for the event would be February 20, 2003. For future work, the framework could be adapted to cover these cases; however, the current performance is acceptable.

Listing 4.7: Example tweet for misclassification of temporal expression extraction.

```
Last Wednesday was 10 years since the Great White fire in Providence (
created at 25.02.2013 19:11)
```

The results show that with our approach, we are able to detect temporal expressions in tweets with high precision. Furthermore, based on these expressions, we are able to calculate a likely date and time when an event took place.

4.2.5.3 Summary

In this section, we presented several studies analyzing the performance of our approaches for named entity and temporal expression recognition. We showed that it is possible to detect named entities in more than 50% of the tweets in our data sets with Spotlight. Furthermore, our approach for location mention detection was able to detect the correct type of location mention with high precision and recall ($P=94.20\%$ and $R=91.29\%$). Furthermore, we showed that we are also able to detect temporal expressions with high precision and recall ($P=92.5\%$ and $R=91.3\%$). In a fourth study, we evaluated the performance of estimating the point in time of an

event mentioned in a tweet and showed that we are able to detect the correct point in time in 89% of the cases.

4.3 Conclusion

In this chapter, we introduced several preprocessing steps needed for creating a system that suffices [R1], [R2], and [R3]. These steps are conducted based on the unstructured data, which was retrieved in the previous step of the processing pipeline. As a result of our approach, we are able to infer temporal information for user-generated content as well as to create structured representation of previously unstructured data.

We presented the following components, which are preliminaries for the contributions of this dissertation:

- We showed which preprocessing steps are conducted to convert unstructured to structured text. With these preprocessing steps, textual content can be prepared in such a way that it can be used for feature generation.
- We presented a set of adaptations applied to standard techniques that allow us to extract named entities and location mentions from unstructured text. We showed how we make use of LOD as a valuable source of background information for various types of entities. Also, we introduced an approach for location mention extraction, which is an important feature for text mining as well as for the geolocalization of social media data. The recognition rate of each individual approach was evaluated on data sets of Twitter messages.
- We proposed an adaptation of a standard approach for extracting temporal expressions and showed how we make use of the results for inferring the temporal dimension of a tweet. This adaptation enables us to calculate a likely point in time when an event occurred, which allows us to determine the temporal dimension of an information item R[3]. We evaluated the accuracy of this approach on a set of tweets and showed that it is capable of determining the correct point in time with 89% of the cases.

In the next chapter, we deal with the question of how to infer the spatial dimension for a tweet R[2].

5 Geolocalization of User-Generated Content

In the last chapter, we dealt with the problem of how to preprocess unstructured text and how to identify named entities and temporal expressions as first automatic preprocessing steps in the framework. As a second step of the automatic preprocessing, we deal with the problem how to infer the spatial dimension of a tweet (R[2]). As a result of this step, spatial information is extracted, which can be used for event clustering (see Figure 7).

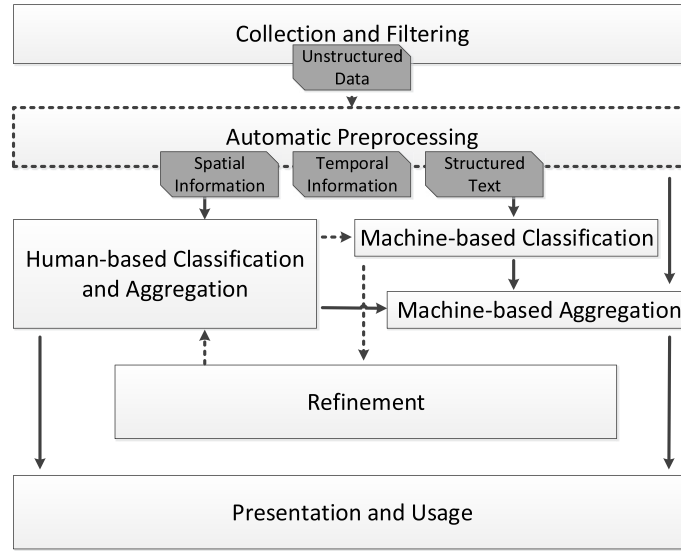


Figure 7.: Automatic preprocessing as a second step of the overall pipeline.

According to requirement R[2], the spatial dimension of each information item needs to be inferred. This dimension is important as determining the location where an incident happened enables a decision maker to relate the incoming information to a specific location. However, according to recent analyses [92], only around 1% of all tweets are explicitly geotagged. Thus, without a possibility to predict the location of tweets, 99% of all tweets cannot be used for estimating the location of an event. Thus, we need to deal with the following question:

- *How can social media data be geolocalized?*

In this chapter, we present a novel approach for the geolocalization of tweets R[2]. This preprocessing step is conducted on the unprocessed text retrieved from the collection step in the framework.

The contributions of this chapter are the following:

- We identify parts of tweets and their metadata suitable for geolocalization.
- We propose a novel approach for the geolocalization of tweets that is capable of inferring the home location of a Twitter user, the point of origin where a tweet was sent, as well as for inferring the location focus of a tweet message.
- We validate the accuracy of our approach on 927K tweets and show that the approach is able to locate 92% of all tweets with a median accuracy of below 30 km. Furthermore, it predicts the user's residence with a median accuracy of below 5.1 km. Finally, the same approach is able to estimate the focus of incident-related tweets within a median accuracy of below 250 m.

The presented geolocalization approach is a general approach; thus, it can easily be applied to different types of user-generated content. However, we tailor the framework to handle specific input data from Twitter.

The section is structured as follows: First, a background on the geolocalization of tweets and a detailed overview of spatial concepts present in Twitter messages and their metadata are provided (see Section 5.1). In Section 5.2, we give an overview of related work. The third section shows our approach (see Section 5.3), followed by an evaluation in Section 5.4. The results are finally discussed in Section 5.5.

Parts of this chapter appeared in [200].

5.1 Background

Only around 1% of all tweets are explicitly geotagged. Hence, approaches for the geolocalization of tweets are needed that infer locations without having explicit geotags. However, simple approaches to determine the location of a tweet are not applicable: The location cannot be estimated using the IP address of a user's device as neither Twitter nor the telecommunication providers allow access to that information for application programmers. Twitter's Search API, which provides spatial filters, relies solely on user profiles, which are often incomplete and incorrect [95]. Extracting location information by other means is challenging as for the geolocalization of tweets, location mentions have to be extracted and mapped to geocoordinates. Several problems arise related to the topic of toponym resolution, which is presented in Section 5.1.1. Furthermore, location mentions are not solely provided in the tweet message, but also in the metadata of the user profile. We give an overview of these mentions in Section 5.1.2.

5.1.1 Toponym Resolution

For the geolocalization of tweets, location mentions have to be identified in the tweet message and in the tweet’s metadata. We rely solely on proper location mentions as common location mentions might lead to heavily biased results if limited contextual information is present to infer the corresponding proper location mention.

As outlined before, detecting proper location mentions (i.e., toponyms) in unstructured texts is difficult due to the characteristics of short texts. For instance, people use different abbreviated named entity mentions for locations (e.g., "LA", "L.A.", or "City of Los Angeles"). Thus, methods that enable the extraction of different toponyms are needed.

Beside the task of identifying these toponyms, we have to deal with mapping the toponyms to geocoordinates. For mapping toponyms to accurate locations, two general problems have to be solved [139]:

1. First, a toponym can refer to multiple geographic locations. For example, *Paris* is referring to 23 cities in the United States. This problem is called the Geo/Geo disambiguation problem.
2. Second, a toponym can relate to entries that can refer to a spatial location and also to a person or a thing. For example, *Vienna* may refer to a city as well as to a person. *As* is used as an adverb but may also refer to a city in Belgium; or *Metro* may reference a city in Indonesia, a train, or a company. This problem is called the Geo/Non-geo disambiguation problem.

Both disambiguation problems have been researched in the area of *toponym resolution* [132] on large text documents. However, both problems are also major challenges when dealing with location information in tweets. In this case, disambiguation is much more complicated as less contextual information is present in the text itself. Still, valuable information is present in the tweet’s metadata.

5.1.2 Spatial Indicators in Tweets

For resolving ambiguous toponyms, it is helpful to leverage different indicators. For example, for distinguishing the European country *Norway* from the equally named city in Australia, the time zone may be a helpful additional indicator. Thus, we make use of various spatial indicators present in the tweet’s metadata. A spatial indicator is defined as follows:

Definition II.5. A *spatial indicator* is a noun or a noun phrase of textual information that helps locate a tweet.

In Figure 8, an example of spatial indicators in a tweet and of a Twitter user profile is shown. Twitter users provide many spatial indicators in their messages and in their profiles. The message text, user account information, website links, current time zone, a dedicated location field, and sometimes even accurate GPS coordinates determined by the user’s mobile device may all be part of a tweet.



Figure 8.: Example for spatial indicators in tweets and user profiles.

In the following, we give an overview of spatial indicators that are present in tweets.

Spatial Indicators in the Tweet Message

The text message of a tweet is at most 140 characters long, is unstructured, and is often written in nonstandard language. Extracting location information from the message is difficult as proper place-names are seldom used, while abbreviations and nicknames are more common. Furthermore, the toponyms may or may not refer to the user’s current location as he could write about a place he is on the way to or where he would like to be. A tweet might even include more than one location in the text (e.g., ”I’d love go to Hawaii or Mauritius”).

Links included in tweets might reference geotagged pictures on Flickr or other links from location-based services. For example, Foursquare allows to “check-in” at a venue resulting in the creation of a tweet with accurate location information. In our data set, we found links to many location-based services such as Foursquare¹⁹ and UberSocial²⁰. As these location-based services are commonly used to inform about the user’s current location, information present on these linked web pages can be used as spatial indicators.

Spatial Indicators in a User’s Profile Information

Twitter users can maintain a personal profile. Furthermore, Twitter adds further information about the user and the tweet. All this information is available as metadata for each tweet, in particular:

Location Field: Users can specify their home location(s) in the location field. The entries in the location field are heterogeneous; the user may, for example, provide

¹⁹ <http://foursquare.com> [Accessed: 19.02.2013]

²⁰ <http://www.ubersocial.com> [Accessed: 15.01.2013]

their home country or their state [80]. Furthermore, abbreviations are commonly used, like *NY* for *New York* or *MN*, which may stand for *Minnesota* but also for the country *Mongolia*. Most of these location entries have a relatively large geographic scope, like *California* or *UK*. Besides real location information, the location field is also used for sarcastic comments or fake location information like *Middleearth* (which is an actual city, but mostly used as a place described in a fantasy book).

Hecht et al. [95], who did the first in-depth study of the location field, showed that only 66% of the entered information has valid geographic information. Furthermore, they showed the reflection of current trends such as the mentioning of Justin Bieber in the text of the location field (e.g. "Biebertown"). Also, 2.6% of the users enter multiple locations. GPS coordinates are part of the location field too, either in decimal (latitude/longitude pairs, for example, "iPhone: 48.1565, 11.5021") or DMS (degrees/minutes/seconds, for example, N 51° 27' 0'' / E 6° 34' 0'') notation. Mostly, these GPS coordinates are provided by mobile devices or mobile applications. Besides correct coordinates, there are also parts of coordinates or IP addresses that could prevent easy parsing.

Websites: In their profiles, Twitter users may provide links to web pages that may, for example, contain personal information. People provide links to Twitter, Facebook, or other social network pages as well as personal websites. Both the website's country code and the website's geocoded IP address are spatial indicators.

Time Zone: The time zone entries in the user's profile describe a region on earth that has a uniform standard time. The time zone is initially set by Twitter and can manually be adjusted by the user. It is typically represented by a city, which is often the capital city of the user's home country (e.g., *London*). On the other hand, the time zone can also describe a larger region without an explicit capital mentioned. For example, the entry "Eastern Time (USA&Canada)" comprises cities such as Montreal, New York City, or Washington DC (see Figure 9).



Figure 9.: Countries covered by time zone entry "Eastern Time (USA&Canada)". (Picture adapted based on [228])

UTC24-Offset: UTC is the time standard used for many World Wide Web standards. Twenty-four main time zones on earth are computed as an offset from UTC, each time zone boundary being 15 degrees of longitude in width, with local variations.

Therefore, the UTC offset is just a means for differentiating a location by longitude compared with the much more precise time zone information (see Figure 10).



Figure 10.: Countries covered by UTC offset of "UTC-05:00". (Picture adapted based on [228])

Coordinates and Place: Depending on the privacy settings, tweets may also contain location information as latitude/longitude coordinate pairs. The coordinates are set when the user sends a tweet from a device with enabled GPS. These device locations are difficult to change and manipulated and can be seen as a very good approximation of the user's position when sending a tweet. Furthermore, Twitter provides an approximate location specified as a bounding box. To create this bounding box, Twitter uses the user's IP address to create the approximation.

5.2 Related Work

Identifying what the geographical location digital content is referring to is a field of extensive research. There are methods to identify the geographic location of digital text documents [216], web pages [60, 15, 257], blogs and news pages [139], and Flickr tags [214, 207, 175].

In the last years, research also dealt with geolocating tweets. The works focused on Twitter can be differentiated along three aspects: the spatial indicators used, the techniques applied, and the focus of geolocalization.

5.2.1 The Use of Spatial Indicators in Related Work

Different information sources are used for geolocalization purposes. The message text is used most of the times; for instance, the approaches proposed in [45, 66, 95, 121, 144, 43] use language models based on the terms in the tweet message. Chandra, Khan, and Muhaya [42] followed these approaches but also took the relationships of the users into account.

Gelernter and Mushegian [80], Sultanik and Fink [222], and Paradesi [169] proposed to apply NER to annotate tweet messages and preprocessing to handle the disambiguation problem. Ikawa, Enoki, and Tatsubori [106] also recommended us-

Table 5.: Overview of related approaches. Spatial indicators and techniques marked with (X) were used by the respective approaches for creating baselines or were part of the background analysis.

	Spatial Indicators				Techniques		Localization Focus		
	Text	Location Field	Social Network	Other	NLP	Gazetteer	Home Loc.	Place of Origin	Message Focus
Eisenstein et al. [66]	X				X		X		
Hecht et al. [95]	X	(X)			X	X	X		
Mahmud et al. [144]	X				X	X	X		
Cheng et al. [45]	X	(X)			X	X	X		
Chandra et al. [42]	X		X		X		X		
Chang et al. [43]	X	(X)			X	X	X		
Gelernter and Mushegian [80]	X				X				X
Sultanik and Fink [222]	X				(X)	X			X
Ikawa et al. [106]	X				X			X	
Kinsella and Murdock [121]	X	(X)			X	(X)	X	X	
Hong et al. [100]	X				X			X	
Ahmed et al. [7]	X				X			X	
Paradesi [169]	X					X		X	
Hale et al. [92]	X	X		(X)		X		X	
Kulshrestha et al. [124]		X		X		X	X		
Abrol and Khan [3]	(X)	(X)	X		X	(X)	X		
Takhteyev et al. [224]		(X)	X		X	(X)	X		
Clodoveu et al. [48]			X			(X)	X		
Mcgee et al. [152]		(X)	X			(X)	X		
Gonzalez et al. [84]			X			(X)	X		
Sadilek et al. [190]			X	(X)	X		X		
Rodrigues et al. [187]			X	(X)	X			X	
Wang et al. [239]	X		X	X	X		X		
Jurgens [115]	X		X	(X)		X	X		
Krishnamurthy and Arlitt [123]				X	-	-	X		
Bouillot et al. [32]	X	X		X		X		X	
MacEachren et al. [143]	X	X		X		X		X	
Han et al. [94]	X	X		X	X		X		
Our Approach	X	X		X		X	X	X	X

ing a language model, but in their approach, they analyzed only keywords from messages created by location-based services like Foursquare.

The algorithm developed by Hong et al. [100] is based on the words a user uses in his tweets. They showed the advantage of identifying topical patterns for geographical regions. Recently, Ahmed, Hong, and Smola [7] extended this approach using a more precise statistical model. Hale et al. [92] analyzed if the language of the message text can be used for geolocalization. They concluded that the language is not an appropriate indicator.

Besides the message, the location field is used for location estimation. Hecht et al. [95] provided an in-depth analysis of the location field. As a result, they concluded that the location field alone does not provide enough information for geolocalization. Hale and Gaffney [92] and Kulshrestha et al. [124] analyzed different geocoders for identifying the location where a user is tweeting from based on the location field.

Instead of the directly usable information of the message or the location field, the relationships of the users are also useful for geolocalization. Abrol and Khan [3] tried to identify the location of a user based on his/her social activities. Takhteyev, Gruz, and Wellman [224]; Clodoveu et al. [48]; and McGee, Caverlee, and Cheng [152] analyzed the relationship between a pair of users and the distance between the pair. Gonzalez et al. [84] focused on the follower relationship and reported that in countries like Brazil, there is a high intracountry locality among users, while in English-speaking countries, the external locality effect is higher. The approach of Sadilek, Kautz, and Bigham [190] is also based on the relationship between users, but in this case, the GPS tags are also used for location inference. Rodrigues et al. [187] also proposed to use the social network in combination with information derived from the tweet message to infer the place of origin for three Brazilian cities. Wang et al. [239] followed a similar approach, but in contrast to previous works, they focused on a Chinese tweet corpus. Recently, Jurgens [115] also proposed to use the social network in combination with location information from other social networks.

Krishnamurthy and Arlitt [123] proposed to use the UTC offset information to get a user's local time and thereby an approximate longitude. They compared their results with the top-level domains of the URL of a user. The results show that users with a URL in the .com domain are distributed around the world, while the rest of the UTC data is lined up with the domain information.

Several approaches tried to combine different information sources. Bouillot, Poncellet, and Roche [32] proposed an approach based on different aspects of user information, like the message, the location field, as well as the language for homonym differentiation. MacEachren et al. [143] developed an application that leverages the geocoded location field, the time zone, the hashtags, and the named entities from the tweet for the geolocalization and geovisual analytics of tweets in crisis manage-

ment. In none of these works, quantitative evaluation results for geolocalization were provided. Most recently, Han, Cook, and Baldwin [93, 94] proposed to combine metadata in the user profile as well as information in the tweet message for location inference. Compared with other approaches, they built a set of statistical classifiers and combined them with each other.

5.2.2 Techniques Used in Related Work

The approaches can be divided into methods mainly based on Natural Language Processing (NLP) that do not use external information and approaches based on geographical dictionaries (gazetteers). The NLP-based approaches were especially designed to estimate the location using language models and context information about the user. Also, statistical models were trained based on the words used in the message.

Opposed to this, the gazetteer approaches used geocoders to determine the place being referred to. These approaches cannot find information that is not present in the gazetteer but have the advantage that no model has to be trained. However, gazetteers were also used several times by the NLP-based approaches on the location field for creating a baseline or training the models.

5.2.3 Focus of Geolocalization in Related Work

All analyzed approaches pursue different goals:

- The **Home Location** is the residence of the user.
- The **Place of Origin** is the location where a tweet was sent.
- The **Message Focus** refers to the location of what the user is tweeting about.

This differentiation is clearly necessary depending on the use case. For incident detection, it is relevant what place a message refers to or where a tweet was sent. For location-based services, the location where a tweet was created is relevant. And for market research, one rather focuses on the user's home location.

5.2.4 Discussion of Related Work

Table 5 provides an overview on the related approaches. Language models are used for the geolocalization of the user as well as for the geolocalization of the tweet. On the other side, the social network is only used for predicting the home location. Except the language model of Kinsella and Murdock [121], there is no approach for detecting both the home location and the location where the tweet was sent. The

advantage of using different information sources at once (e.g., language information as well as place-names from the location field and the message) was shown several times by [32, 143, 95].

Our method is innovative in several aspects compared with related work. Our multi-indicator approach uses a variety of spatial indicators to solve the geolocalization problem. We are able to determine the point of origin of the tweet, the home location of a user, as well as the message focus by taking the message, the location field, and further metadata into account.

5.3 Approach

In this section, we present our approach for the geolocalization of tweets. In order to estimate the point of origin for a tweet, we use a variety of spatial indicators. In the first section (see Section 5.3.1), we present how spatial indicators are combined to form a single geolocation estimate for determining the place of origin of a tweet and the home location of a Twitter user. In Section 5.3.2, we present our approach for estimating the message focus of a tweet.

5.3.1 Approach for Determining the Place of Origin of a Tweet and the Home Location of a Twitter User

In this section, we present the general approach for the geolocalization of tweets and show how it is applied for determining the place of origin of a tweet and the home location of a Twitter user. Our proposed method consists of four steps:

1. **Detection of Spatial Indicators:** Spatial indicators are location information that allows geolocalization. Our method spots spatial indicators in the text message and in the user profile of a Twitter user.
2. **Polygon Mapping:** Each spatial indicator refers to (at least) one geographical area. We determine that area and represent it with a polygon.
3. **Polygon Height:** As some spatial indicators are more reliable than others, we attribute a variable height to each polygon. The height is computed based on weights determined using an optimization algorithm and the reported uncertainty of the spatial indicator for the currently analyzed case.
4. **Polygon Stacking:** By intersecting and stacking the 3-D polygons over each other, a height map is built. The highest area in this height map is then used for geolocalization.

Properties of spatial indicators: Usually one or a multitude of spatial indicators can be extracted from a single tweet using different resolution methods. In order to

successfully combine the spatial indicators, it is necessary to understand their basic properties:

- **Contradiction:** The spatial indicators extracted from a tweet can coincide (e.g., location field: Paris, message: *Nice weather in Paris*), or they can be contradictory (e.g., location field: *Paris*, message: *Nice weather in Athens*).
- **Scale:** The spatial indicators can relate to areas of different scale. Consider, for instance, the spatial indicators extracted from the location mentions "France" and "Eiffel Tower" that may occur together in a Twitter message and which represent geographical areas of vastly different size.
- **Ambiguity:** As discussed above, spatial indicators are ambiguous, such as the different cities called *Paris*. Further ambiguity may be a result of spelling errors, the use of abbreviations, incomplete information, and slang. Gazetteers usually provide a list of different geographical interpretations of a geographical name with ratings of their uncertainty (e.g., based on the edit distance of a misspelled city name).

Polygon mapping: Simple solutions to combine the spatial indicators into a single location estimate, such as computing the average of the coordinates given by each spatial indicator, are bound to fail due to problems with contradiction, scale, and ambiguity. In order to get a good combined estimate, we adopted the approach of Woodruff and Plaunt [246] for localizing bibliographical text documents, which is based on intersecting the geographical outlines of the geographical areas that the spatial indicators refer to. These geographical outlines are represented by polygons. The mapping from spatial indicators to polygons is either done directly by the resolution method itself or indirectly using coordinate pairs that are provided by the resolution method and mapping to an appropriate surrounding area in a spatial database (see below).

Polygon height: To arrive at a uniform prediction, a height is attributed to each polygon, making it a three-dimensional shape. The height allows for modeling the uncertainty that may come with a spatial indicator. This uncertainty can be an outcome from the method itself, which may sometimes make wrong predictions, or from the inherent inaccuracy of a spatial indicator (e.g., the time zone indicator). Therefore, the final polygon height is determined based on two factors: First, it is based on the quality of the resolution method that was used. Based on our evaluation results, we assign a quality factor Q_{ext} to each method based on how well it contributes to predict the tweet's location. The value Q_{ext} is determined using the simplex method of Nelder and Mead (see below). In addition to this "external" quality measure, many methods also provide an internal assessment of the quality when more than one alternative is suggested.

The internal quality measure $Q_{\text{int}}(x)$ provides an estimation for the quality of the x -th alternative. Since different resolution methods have vastly different scales when reporting this internal quality measure, they are normalized to a $[0, 1]$ interval. Resolution methods returning only one result are rated with $Q_{\text{int}} = 1$. The height h of the polygon representing the x -th alternative is then computed as $h(x) = Q_{\text{ext}} \cdot Q_{\text{int}}(x)$. The rationale behind this is that the normalized internal quality measures can be weighted using our external quality measures.

Polygon stacking: Once all three-dimensional polygon shapes are determined, they are stacked one over the other and form a height profile (see Figure 11). The highest area in that height profile is then found and its polygon outline is determined as the intersection of the contributing polygons. The geolocation is estimated as the geometric center of that area as a coordinate pair.

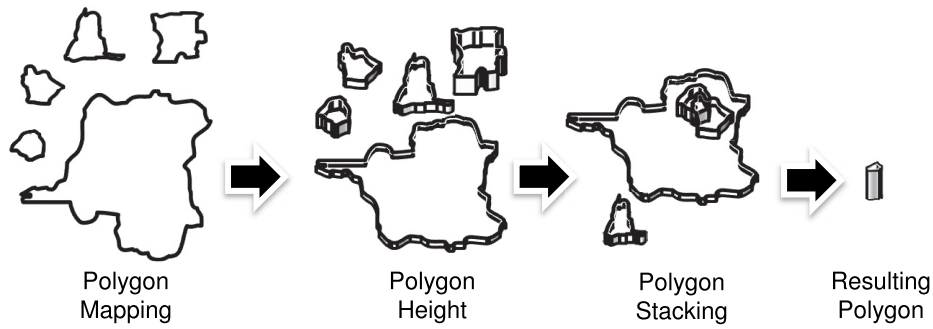


Figure 11.: The height profile is determined by stacking the three-dimensional polygon shapes over each other.

Figure 12 illustrates how an example tweet consisting of several spatial indicators is processed using our approach. As a result of this process, we estimate the location of the tweet with a confidence value.

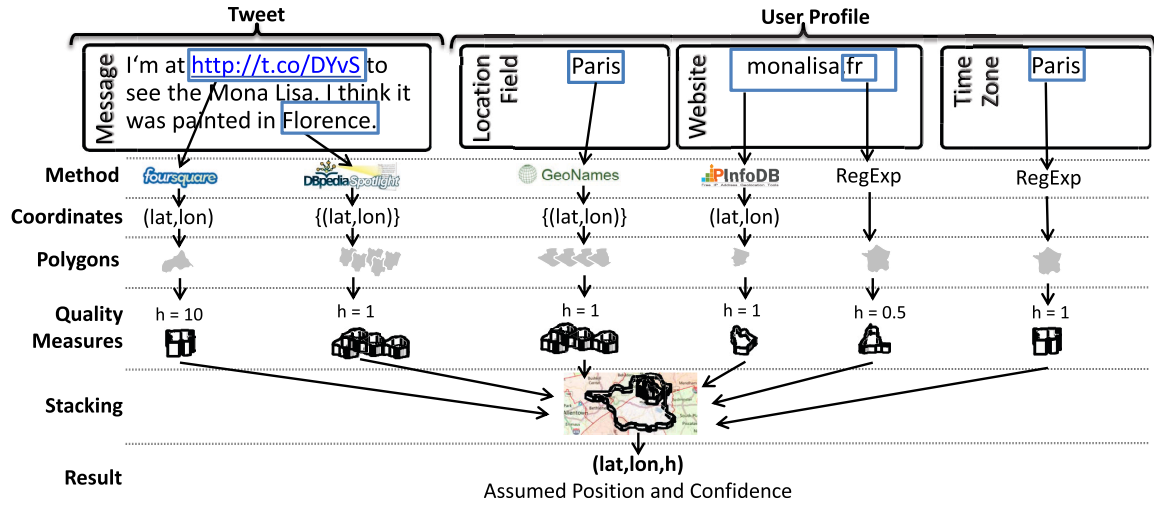


Figure 12.: Example pipeline for our approach: Spatial indicators are identified based on the methods described. The results are either a pair of coordinates (lat, lon) or a set of coordinates and quality measures. The coordinates are mapped to the corresponding polygons. Then the external quality measures are applied before conducting the stacking. As a result we estimate the location of the tweet with a confidence value.

Resolution Methods for Extracting Spatial Indicators

Hereinafter, we describe how the different spatial indicators described in Section 5.1 are extracted from tweets. Depending on the size of the identified spatial entity, extracted coordinates are either used directly or are mapped to polygons.

Tweet Message: From the text, we extract named entities using DBpedia Spotlight, a NER service that identifies entities in texts and maps them to DBpedia entities (**SP**). For example, in the tweet "yeah, watching muse at fedex field!!!", the text *muse* is recognized as a named entity and mapped to [http://dbpedia.org/resource/Muse_\(Band\)](http://dbpedia.org/resource/Muse_(Band)), as well as *fedex field* is mapped to http://dbpedia.org/resource/FedEx_Field. If the extracted named entities are location mentions, these geographic entities (such as *FedEx Field*) have coordinates in DBpedia. We extract and use those coordinates for polygon mapping. Entities without coordinates (such as *Muse*) are discarded. For calculating $Q_{int}(x)$, we use the confidence values provided by Spotlight.

For processing information from location-based services (**LBS**), we analyzed our data set for the occurrences of the most common services. In this case, we extract coor-

ordinates from UberSocial²¹, TrendsMap²², Flickr²³, Rocketatchi²⁴, and Foursquare²⁵ based on the information provided on the web page. For every location-based service, we identify the coordinates based on predefined patterns. For instance, for extracting geolocations for Foursquare check-ins, we use the meta tags referencing the venues and corresponding location information.

Location Field: For toponym resolution in the location field, we use GeoNames²⁶ (GN). GeoNames is a gazetteer that contains more than 10 million entries about geographic entities in different languages. This includes countries, cities, as well as building and street names. Using the full text search, GeoNames returns a list of possible results with a confidence score, which we use for calculating $Q_{\text{int}}(x)$. As GeoNames is not able to resolve all location field entries directly, we preprocess the entries in different steps if no results are returned:

- GeoNames has problems processing unaligned text segments. We solve this by text preprocessing (GN-1).
- We extract several toponyms from the location entry (GN-2). First, as a lot of the location field entries contain separators like "|" (e.g., "Salvador | Bahia | Brasil"), we split this entry into a list of entities. Furthermore, more general location information is often provided in parentheses (e.g., "Berlin, Germany (Europe)"). In this case, we extract the content of the parentheses and try to resolve the first comma group and the parentheses itself in GeoNames.
- As gazetteers often have problems with city-level entities like local places and their nicknames, we use DBpedia Spotlight to annotate the entry in the location field (GN-3). In this case, commonly used nicknames like "The Big Apple" can be retrieved.
- As a last means for extracting toponyms, we split the whole location entry into a list of words (GN-4). Every word is then sent to GeoNames.

As previously mentioned, coordinates are also part of the location field. For extracting these, we use regular expressions to identify them in decimal or the DMS notation (COD). As location-based services do not follow a common pattern for setting coordinate entries, regular expressions were adapted to match most of the common cases. For instance, analyzing entries of location fields in DMS notation shows that numbers are set before the cardinal direction as well as after.

²¹ <http://www.ubersocial.com> [Accessed: 19.02.2013]

²² <http://trendsmat.com> [Accessed: 19.02.2013]

²³ <https://www.flickr.com/> [Accessed: 19.02.2013]

²⁴ <http://socialnetworkingappsearch.com/rocketatchi-app-for-iphone-ipad-387274897/> [Accessed: 19.02.2013]

²⁵ <https://foursquare.com/> [Accessed: 19.02.2013]

²⁶ <http://www.GeoNames.org/> [Accessed: 19.02.2013]

Website: To handle the website metadata field, we follow a twofold approach. First, we extract the top-level domain using a regular expression (**WS-1**). The top-level domains are then matched against country codes using a manually created mapping of country codes and the corresponding country names. *.com*, *.net*, *.org* are not processed in this case, as they do not provide any helpful location information [123]. To provide estimations for these cases, we also extract the IP addresses using the host names (**WS-2**). Coordinates are then retrieved using IPinfoDB²⁷. The API is an IP geolocation lookup service that provides coordinates for a domain name or an IP address.

Time Zone: Our analysis of the different time zone entries has shown that these are mostly provided in a standardized format stating the capital of the home country. Besides these kinds of entries, United States and Canadian time zone entries are also present like *Central Time (USA&Canada)*. The provided time zone entries can be used directly as they are machine generated (**TZ**).

Mapping to Polygons

To enable the mapping of geocoordinates to polygons, we built a spatial database with polygons suitable for every spatial indicator.

Tweet Message and Location Field: For mapping the coordinates retrieved from the message and the location field, we use polygons of the world's administrative areas. For example, "the Bronx" can be retrieved as part of the administrative districts of New York City, allowing us to narrow down our estimation as good as possible. The polygons used for this were retrieved from the GADM database of Global Administrative Areas.²⁸ For mapping coordinates retrieved from location-based services, we use a circle of 100 m radius around the position.

Website: As the website entries might relate to the home country of the user, but not the hometown, we use country polygons for mapping the website entries. In this case, the polygons are retrieved from ThematicMapping.²⁹ The extracted country names from the top-level domains are then matched to the polygons representing the world borders.

Time Zone: For mapping the time zones, we use polygons retrieved from the IANA Time Zone Database.³⁰ In this case, the polygons for the time zones of the United States, Canada, Russia, and China have been aggregated manually as they are not present in the initial data set. Furthermore, the polygons for the time zones spanning multiple countries like the *Central Standard Time (CST)* or *Pacific Standard Time*

²⁷ <http://ipinfodb.com/> [Accessed: 01.03.2014]

²⁸ <http://www.gadm.org> [Accessed: 11.01.2013]

²⁹ http://thematicmapping.org/downloads/world_borders.php [Accessed: 11.01.2013]

³⁰ <http://efele.net/maps/tz/world/> [Accessed: 11.01.2013]

(PST) were created manually based on the regions contained in the corresponding time zone.

Determining External Quality Measures

As stated before, we assign a quality factor Q_{ext} to each method based on how well it contributes to predict the tweet's location. For instance, the time zone is a very imprecise estimator for a geolocation, whereas information from location-based services is very precise up to building level. Since not all of our indicators are expected to perform equally well, we assign a quality measure to each method based on how well it contributes to predict the tweet's location. To determine a good external quality measure of each approach, we first defined the problem of optimizing the quality measures as an unconstrained minimization problem:

$$\text{Minimize } F(x) = \sqrt{\frac{\sum_{t \in T} d(l_{\text{act}}(t_n), l_{\text{est}}(t_n))^2}{|T|}} \quad (5)$$

We apply this approach to determine the external quality measures for estimating the location where a tweet was sent. The objective function is the Root Mean Squared Error (see Section 5.4) of the sum of all error distances, calculated as distances d between the ground truth $l_{\text{act}}(t_n)$ and the estimated geolocation $l_{\text{est}}(t_n)$ for a set of tweets $T = (t_1, t_2, \dots, t_n)$. The device locations are used as ground truth in this calculation.

For finding a local optimum for this problem, we use the downhill simplex method of Nelder and Mead [160]. To apply the method, we regard the weight of each method as a variable of our objective function. A vector of zeros for all feature weights is used as an initial guess. With the optimization method, we are able to calculate a local optimum for minimizing the objective function. The results of these optimization steps are presented and discussed in Section 5.4.

5.3.2 Approach for Estimating the Focus of Incident-Related Tweets

For small-scale incidents, we cannot necessarily assume that the tweeter is close to the incident. For example, someone is in a traffic jam with a length of several kilometers. Thus, we also have to take the message focus into account. Furthermore, for small-scale incidents street-level precision is needed to determine the exact location of an incident. In the following, we present our approach for estimating the focus of incident-related tweets.

The approach is based on the results of the first approach and uses the location mention extraction approach presented in Section 4.2. Thus, it is highly optimized for a

city for which the geolocalization shall take place compared with the first approach, which was not optimized for a specific location.

The following incident-related tweet shown in Listing 5.1 was created in Seattle and is an example that shows how location mentions are used.

Listing 5.1: Incident-related tweet showing usage of location mentions.

```
Collision at Rainier Ave S & S Henderson St 2 eastbound lanes blocked  
w/o Rainier Ave S
```

Compared with the previous approach, the properties of spatial indicators change:

- **Contradiction:** The spatial indicators extracted from an incident-related tweet message can coincide, but they are *not* contradictory.
- **Scale:** The spatial indicators can relate to areas of different scale, but the scale is highly limited.
- **Ambiguity:** Assuming that the city where the tweet was sent is known, spatial indicators have no or limited ambiguity. This is because street names or place-names are unique in each city. However, as before, ambiguity might be a result of spelling errors, the use of abbreviations, incomplete information, and slang.

Before applying the approach, we first infer the country and city where a tweet was sent before determining the message focus. The results are used as input for estimating the message focus. This input is important as street names and other city-level proper location mentions can only be disambiguated if the city is known. The order of the subsequent steps are shown in Figure 13.

Polygon mapping: As a first step for our approach, we use the location mention extraction approach presented in Section 4.2.3 to detect possible location mentions in the tweet message. Based on this information, we create triples of combinations of consecutive words (word-n-grams) to determine likely location names. For instance, for the location mention "S Henderson St", the n-grams "Henderson", "S Henderson", "Henderson S", etc., are created. We then use the MapQuest Nominatim API³¹ to map each n-gram to a location in the corresponding city. This results in several sets of coordinate pairs for each n-gram. Based on these pairs, we create a polygon. In this case, no polygon database is used. However, we use the corresponding coordinate pairs returned by the APIs to create a polygon of the place. As a last step, we remove redundant polygons as some n-grams refer to the same location.

Polygon height: In this approach, only the quality of the resolution method is used as an internal quality measure $Q_{\text{int}}(x)$. The external quality measure is not used as only one method is used for mapping to coordinates.

³¹ <http://developer.mapquest.com/web/products/open/nominatim> [Accessed: 15.01.2014]

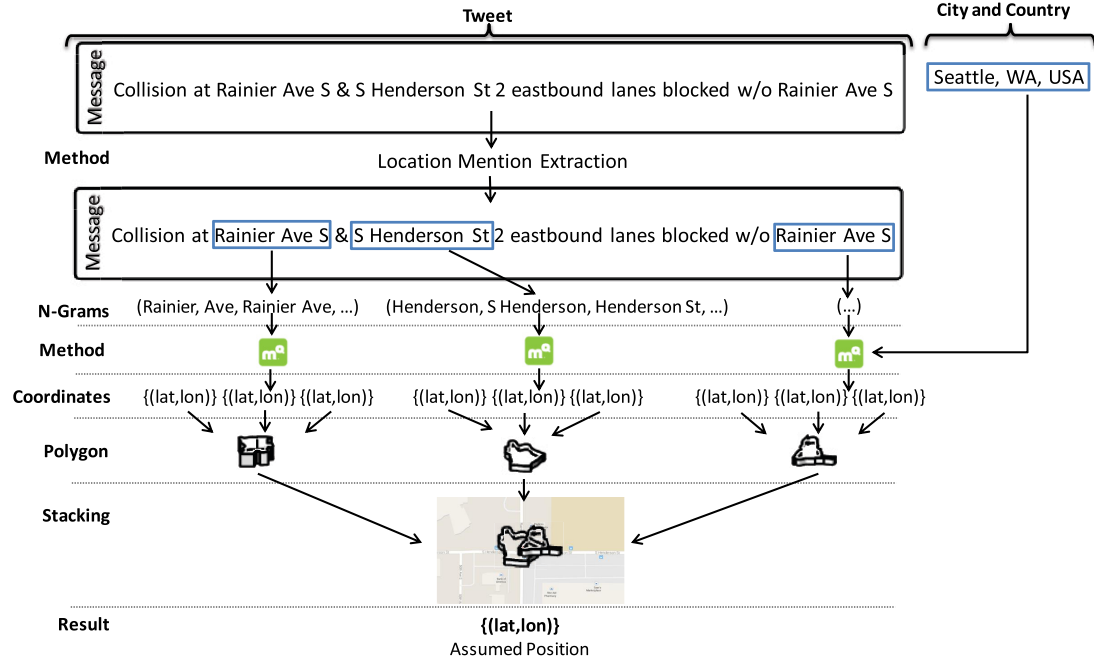


Figure 13.: Example pipeline for our approach for street-level geolocation of incident-related tweets.

Polygon stacking: As presented in the approach before, once all polygons are determined, they are stacked one over the other. The highest area in that height profile is then found, and its polygon outline is determined as the intersection of the contributing polygons. In this case, the polygon is used as estimation of the area.

The presented approach allows the street-level geolocation of the message focus of tweets. However, it has limitations when it comes to identifying lanes and directions. For instance, in the example tweet, "2 eastbound lanes" are mentioned. The presented approach does not take this information into account. Coping with this limitation is open for future work.

5.4 Evaluation

In the following section, we present the evaluation of our methods. After introducing the ground truth data sets and metrics used for our evaluations, we present the result of five studies. First, we show the results of estimating the quality measures. Second, we present the evaluation results for each single spatial indicator for the geolocation of the point of origin of a tweet. Third, we show the performance of the overall approach. Fourth, we also show the performance of the approach for

determining the home location of a user. Fifth, we present the evaluation results for the approach for determining the message focus of incident-related tweets.

5.4.1 Data Set and Metrics

In the following, we introduce the data sets and metrics used for our evaluation.

Data Sets

For developing ground truth data, we used the SET_GEO as an initial data set (see Section 3.2.1). From these tweets, we extracted 1.03 million messages with device locations to use them for our evaluation. No further preprocessing was applied on the sample set to keep it representative for a real-world scenario. For implementing our approach, we used 10% randomly selected tweets from the data set with device locations for tuning the identification of spatial indicators and 90% for testing.

- **SET_PO:** 927.000 tweets with device locations.

For evaluating our approach for estimating the user's residence, we created a smaller sample set based on data set SET_PO. For this, we identified all tweets for users with more than 20 tweets in the data set and manually geocoded the likely residence of 500 randomly selected users. The final set consists of 17.270 tweets.

- **SET_HL:** 17.270 tweets for 500 Twitter users.

For evaluating our approach for determining the message focus of incident-related tweets, we used the SET_CITY_1 and SET_CITY_2 data sets as base (see Section 7.5). We applied the incident keyword filtering presented in Section 3.2.2. From the resulting set, we randomly selected 2,000 tweets containing at least one incident-related keyword. From this data set, we selected the 100 incident-related tweets that contain location mentions.

Each of the tweets was then manually geocoded based on the information provided in the tweet message. This manual coding is rather difficult as, for instance, the tweet "Accident reported in Bothell - SB 405 before SR 527, right lane. Already looks slow from nearly I-5" needs manual identification of the correct geolocation based on all locations mentioned in the tweet. The final data set is as follows:

- **SET_MF:** 100 incident-related tweets.

Metrics:

To evaluate our approaches, we compared the coordinate pair estimation of each approach with a ground truth. For evaluating the performance of each spatial indicator and the combined approach for the geolocalization of a tweet, we use the

device location as ground truth. For determining the user's residence and geolocating incident-related tweets, we used the manually provided geotags as ground truth.

In the following sections, we provide the following error metrics to ensure comparability to related work. The *Error Distance* is the distance in kilometers or miles between the ground truth $l_{act}(t_n)$ and the estimated geolocation $l_{est}(t_n)$ for a set of tweets $T = (t_1, t_2, \dots, t_n)$. The Error Distance is defined in Equation 6 as follows:

$$ErrorDistance(t_n) = d(l_{act}(t_n), l_{est}(t_n)) \quad (6)$$

We further define the *Average Error Distance* as shown in Equation 7:

$$AED(t_n) = \frac{\sum_{t_n \in T} ErrorDistance(t_n)}{|T|} \quad (7)$$

Based on the AED we calculate the *Root Mean Squared Error* as shown in Equation 8.

$$RMSE(t_n) = \sqrt{\frac{\sum_{t_n \in T} ErrorDistance(t_n)^2}{|T|}} \quad (8)$$

As the RMSE is more sensitive to large errors, we also report the *Median* which is calculated for the ordered sample of n error distances:

$$MED(t_n) = \tilde{t}_n = \begin{cases} ErrorDistance_{\frac{n+1}{2}} & \text{if } n \text{ uneven} \\ \frac{1}{2} (ErrorDistance_{\frac{n}{2}} + x_{\frac{n}{2}+1}) & \text{if } n \text{ even.} \end{cases} \quad (9)$$

Finally, we report the *recall*, which is the number of tweets with identified spatial indicators compared with the number of all tweets.

In the following sections, we provide the results for the spatial indicators as defined in Section 5.3:

- **SP:** Extraction of named entities using Spotlight.
- **LBS:** Information from location-based services.
- **TZ:** Time zone entries in the user profile.
- **WS-1, WS-2:** Top-level domain (WS-1) and IP addresses of websites (WS-2).
- **GN-1 to GN-4:** Location field entry processing using GeoNames with different levels of optimization.
- **COD:** Extraction of coordinates from the location field.

5.4.2 Determining External Quality Measures

To calculate the external quality measures, we used a holdout sample set of 10,000 randomly chosen tweets from SET_PO and applied the approach for determining the optimal values presented in Section 5.3. A vector of zeros for all feature weights is used as an initial guess.

Table 6.: Optimal external quality measures for each spatial indicator.

	SP	LBS	TZ	WS-1	WS-2	COD	GN	GN-1	GN-2	GN-3	GN-4
Q_{ext}	0.87	4.26	1.12	1.07	-2.32	2.72	1.51	2.01	1.67	1.96	-0.54

The determined weights used for all studies are shown in Table 6. The external quality measures for precise spatial indicators such as the coordinates and the location-based services indicators are high (LBS=4.26, COD=2.72). The first three GeoNames optimizations provide better results compared with the plain GeoNames approach (GN=1.51, GN-1=2.01, GN-2=1.67, GN-3=1.96). The fourth optimization, which processes every word in the location field, provides worse results (GN-4=-0.54). Processing the time zone as well as the top-level domains is also contributing to the overall result (TZ=1.12, WS-1=1.07) as well as the message processing based on DBpedia Spotlight (SP=0.87). Using the IP addresses does not provide valuable estimations (WS-2=-2.32).

As we show in the evaluations, applying the external quality measures leads to much better prediction results. However, as the downhill simplex method approximates a local optimum, better results could be achieved with a different initial guess or other optimization algorithms. Furthermore, we use the weights for all studies, although weights might change depending on the geolocalization focus.

5.4.3 Study 1: Evaluation of Single Spatial Indicators

In this study, we evaluated the different approaches for every spatial indicator itself before combining the approaches. For evaluation, we used Set SET_PO. The evaluation results are shown in Table 7 and Table 8.

Tweet Message: The method SP for tagging the messages identifies toponyms in 5.13% of the tweets. The overall estimation with a median error distance of 1,100 km is not suitable for location estimation. DBpedia Spotlight retrieves good estimations on messages mentioning the current location as toponyms in the text, which are created by location-based services. Furthermore, @-mentions like '@Bryant Park' provide good estimations. On the other side, DBpedia Spotlight has

Table 7.: Results of the individual indicator approaches (in km) and external quality measures of the indicators.

	SP	LBS	TZ	WS-1	WS-2
RMSE	5939	403	4229	4896	7230
AED	3689	15.41	2600	2618	5529
MED	1100	0.01	1543	494	3287
Recall	5.13%	18.25%	81.22%	6.46%	34.40%
Q_{ext}	0.87	4.26	1.12	1.07	-2.32

some problems with the nonstandard language in tweets, resulting, for example, in regular words being identified as toponyms.

In contrast, using the LBS method, we get a high precision using the links created by location-based services with about 97% within a 1 km radius, which makes this a suitable source for estimations. The recall of the LBS method is rather low with 18.25%. The result for applying the LBS and SP approaches are shown in Figure 14.

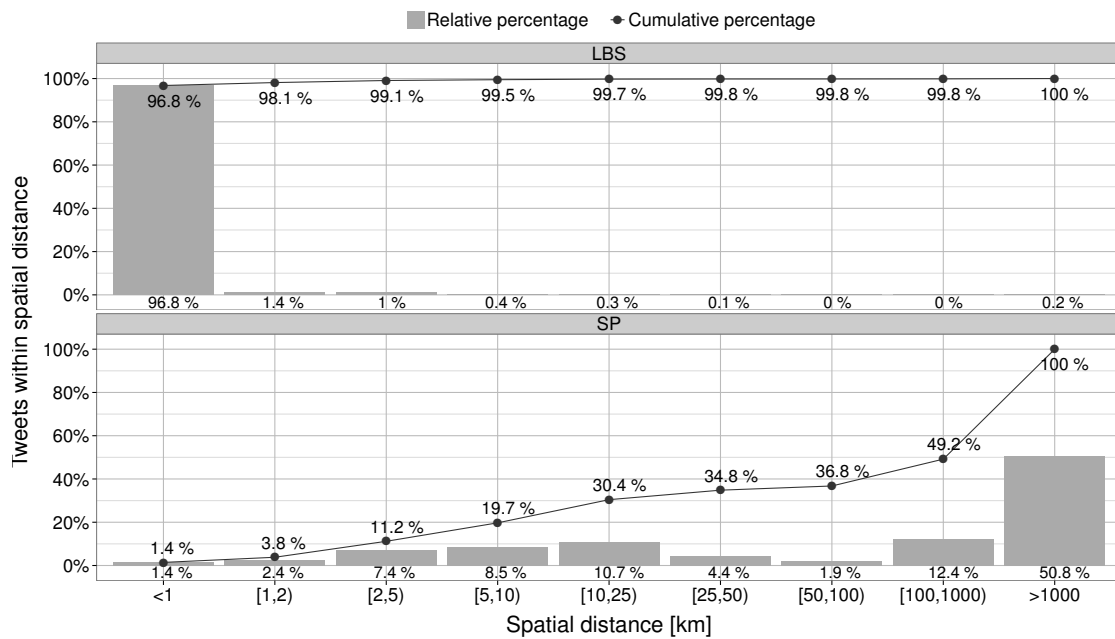


Figure 14.: Overview of evaluation results for LBS and SP. The y-axis of each chart provides the percentages of the tweets within spatial distance. The x-axis shows the spatial distance in km.

Location Field: Using only the coordinates (COD) provided in the location field results in a low recall of 7.73%. The low recall indicates that only a few coordinates are present in the location field. However, the precision of this approach is very

high as 77% are within a 25 km radius and 31% within 10 km. In this case, some outliers result from large differences between the coordinates in the location field entry and the real position. These outliers might be a result of late updates of the device position (e.g., during long-distance flights).

Adding the plain GeoNames approach, GN, we get a good recall of 65.82%. Furthermore, the median error distance of 23.30 km is a result of estimating 62% within a 50 km radius and even 52% within 25 km. Errors are the result of location field entries with multiple toponyms, with parentheses and other combinations that cannot be parsed. With the first optimization GN-1, we can increase the recall by 2% and a concurrent increase of median error distance by 0.7 km. The second optimization GN-2 further increases the recall by 3% without a significant loss of precision. In contrast, the third optimization GN-3 results in a loss of precision while further increasing the recall by 2%. The fourth optimization GN-4 increases the recall to 83% but also reduces the accuracy considerably. It is still possible to estimate 51% of the tweets in a 50 km radius, but the median error distance with more than 1,000 km is much higher.

Table 8.: Results of the individual indicator approaches (in km) and external quality measures of the indicators.

	COD	GN	GN-1	GN-2	GN-3	GN-4
RMSE	1670	3402	3432	3539	3631	4618
AED	349	1354	1320	1380	1459	2188
MED	9.25	23.3	22.65	22.63	25.46	41.40
Recall	7.73%	63.55%	65.82%	69.03%	71.64%	83.29%
Q_{ext}	2.72	1.51	2.01	1.67	1.96	-0.54

As an overall result, extracting toponyms from the location field performs quite well. It still needs more discriminators for better precision (e.g., "loading..." is mapped to *Port Bonython Loading Terminal*). Furthermore, cases where people are on holidays or business trips result in high error rates. For resolving these problems, more information about the user has to be used to identify these cases. Our analysis of the location field further showed that people enter IP addresses, dates, and incomplete coordinates that are not used by our approach. The results for applying the approach on the location field are shown in Figure 15.

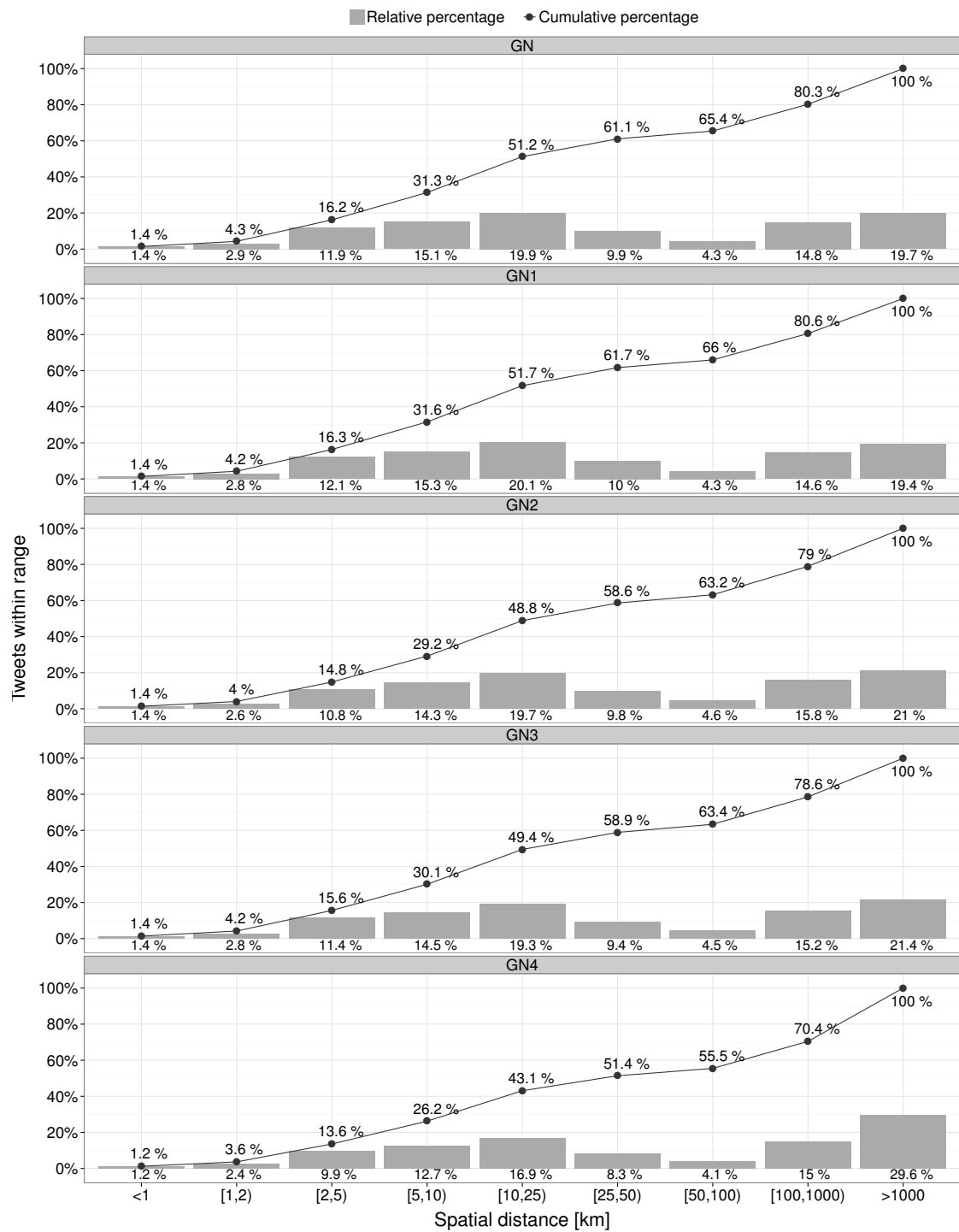


Figure 15.: Overview of evaluation results for applying the approaches on the location field. The y-axis of each chart provides the percentages of the tweets within spatial distance. The x-axis shows the spatial distance in km.

Time Zone and Website: The estimation based on the time zone approach (TZ) is useful for geolocating 81.22% of the tweets. This approach results in a low precision because we use a polygon with the size of the whole country. The same applies for both website handling approaches. The first website approach WS-1 has a low recall of 6.46% because website information is either not provided or related to a top-level domain, which we do not extract. Using the IP addresses in approach WS-2 is also imprecise, but the recall is 34.40% because all websites are used. The precision is even lower than the top-level domain approach. Same as the time zone approach, the two website approaches have low precision because we use the country-wide polygons. All of these approaches are good estimators for smaller countries such as the Netherlands, but loose precision on large countries like the United States. Nevertheless, the provided information can be valuable to differentiate toponyms extracted from the other approaches. The results for applying the approaches are shown in Figure 16.

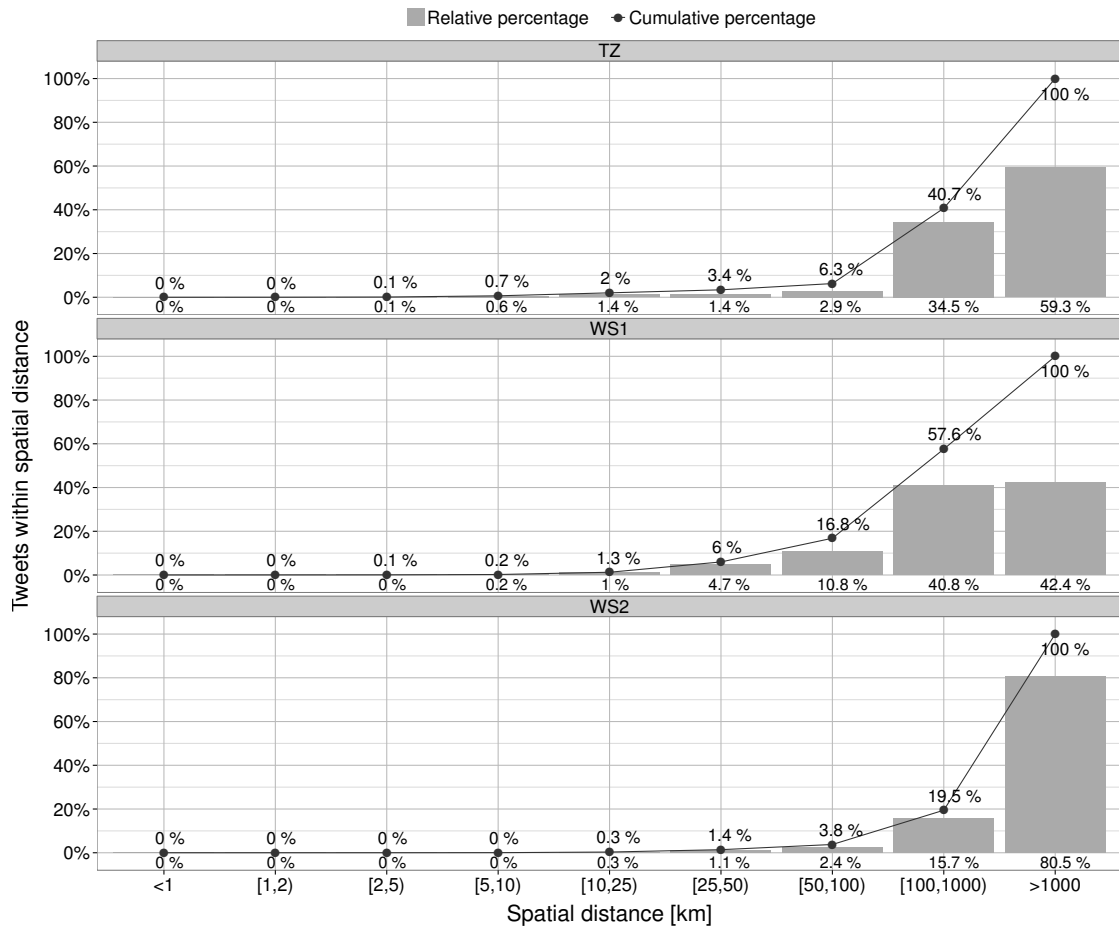


Figure 16.: Overview of evaluation results for TZ, WS-1, and WS-2. The y-axis of each chart provides the percentages of the tweets within spatial distance. The x-axis shows the spatial distance in km.

5.4.4 Study 2: Evaluation of Estimating the Place of Origin

In this section, we present the evaluation of our approach for estimating the place of origin of a tweet. For the evaluation, we discarded the fourth GeoNames optimization GN-4 as well as the approach based on the IP addresses WS-2 because their external quality measures are less than zero (cf. Table 6).

As an overall result, we are able to create estimations for 92% of the tweets in our data set with a median error distance of 29.66 km (cf. Figure 17). We are able to estimate 53% of the tweets within a 50 km radius. The use of optimized quality measures (QM) drastically increases our estimation (cf. Table 9), with a small reduction in recall that results from disposing the two mentioned approaches.

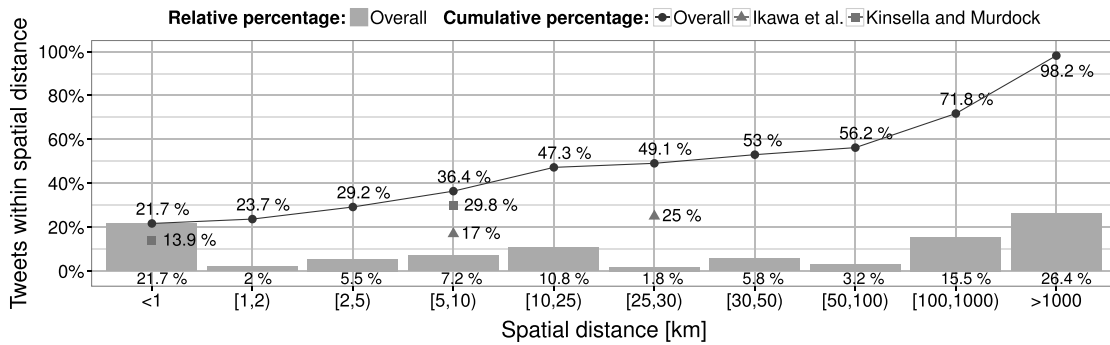


Figure 17.: Overview of the overall results with external quality measures (bottom right). The y-axis of the chart provides the percentages of the tweets within spatial distance. The x-axis shows the spatial distance in km.

Table 9.: Results of the overall geolocalization approach for tweets with and without external quality measures.

	RMSE	AED	MED	Recall
w/o QM	4,159 km	1,931 km	64.46 km	95.10%
with QM	3,310 km	1,408 km	29.66km	92.26%

Ikawa et al. [106] reported a precision of 17% in a 10 km radius and 25% in a 30 km radius. Compared with this, we exceed their results with 37% on 10 km and 48% on 25 km. Kinsella and Murdock [121] reported a precision of 13.9% on "zip code level" and 29.8% on "town level". If we assume a precision of 1 km as zip code level and a precision of 10 km as town level, we also exceed their results as we are able to estimate 22% on a 1 km radius and 37% on a 10 km radius³². We omit

³² We compare with the evaluation based on the worldwide FIREHOSE feed as the Spritzer feed was restricted to 10 towns by Kinsella and Murdock.

a comparison with [169, 92, 187, 7] as they restricted their data sets beforehand to a nonworldwide set. Furthermore, Hong et al. [100], Bouillot et al. [32], and MacEachren et al. [143] do not provide quantitative results. Our results show that we outperform current state-of-the-art tweet geolocalization.

Since our test set consists of those tweets for which we know the coordinates and this selection may be biased, we also tested our approach to detect spatial indicators on a random sample of 10,000 tweets from the whole Spritzer data set. This data set consists of tweets as they are sent every day. No further filtering was applied, thus giving us a general impression of how our approach performs. In this case, no quality measures or mappings to polygons were applied. The results show that our approach would also perform well on a data set with and without device locations (cf. Table 10). Even a suspected decrease of recall in the location-based services indicator could not be found. Though the use of LBS indicators might appear as skewing the results since they are trivial to locate, 22.19% of all tweets in a representative sample are LBS-related tweets; thus, taking that information into account is a valid approach. Also, for only 1% of all location-based services indicators, coordinate entries in the location field are present. Nevertheless, the differences in recall indicate that our approach can be tuned to match yet unknown cases (e.g., previously unknown top-level domains). Furthermore, spatial indicators that are currently not mapped to polygons, which is a result of imprecise location information in the different approaches we apply, could be detected.

Table 10.: Recall of individual indicator approaches on a random and unfiltered sample of the Spritzer stream.

SP	LBS	TZ	WS-1	WS-2	COD
5.66%	22.19%	96.24%	17.43%	79.15%	4.54%
GN	GN-1	GN-2	GN-3	GN-4	
68.67%	70.16%	72.51%	74.74%	82.32%	

5.4.5 Study 3: Evaluation of Estimating the Home Location

To show the applicability of our approach for estimating the user’s residence, we evaluated it on SET_HL. For estimating the quality of our approach, we compared the ground truth geocodes with our estimations. The estimations were created based on the spatial indicators extracted from all tweets of a user, which is different compared with the geolocalization of a tweet as this approach uses only spatial indicators of one tweet. The evaluation results are shown in Table 18 and Figure 18, respectively.

Table 11.: Results of the overall approach for estimating the user’s residence.

	RMSE	AED	MED	Recall
with QM	2,281 km	751 km	5.05 km	100%

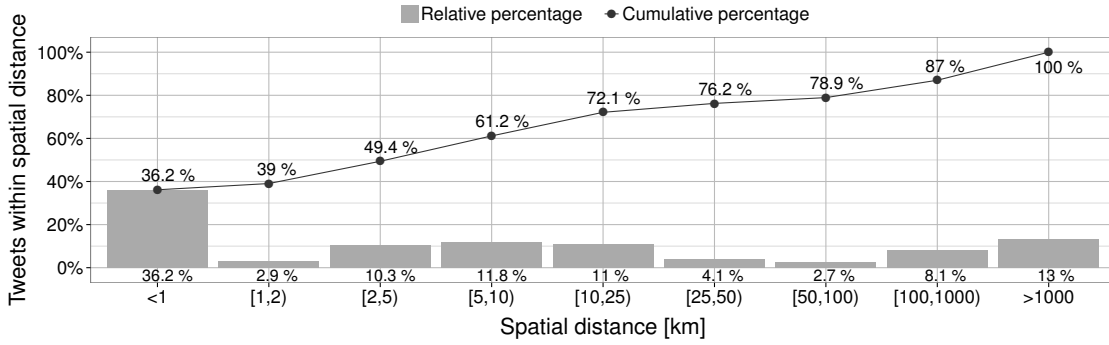


Figure 18.: Overview of the overall results for estimating the user’s residence. The y-axis of the chart provides the percentages of the tweets within spatial distance. The x-axis shows the spatial distance in km.

In Table 12, we provide evaluation results of related approaches. Though all approaches used different ground truth data sets, all approaches used data sets with more than 500 users; thus, the results are comparable to some extent. We omit a comparison with [239] as they do not provide results with respect to estimation distances.

In the most relevant work from [144], the authors reported that their approach is able to estimate the user’s residence in 62% of the cases in a 100 mi radius. With our approach, we are able to estimate 79% of the user’s residences in a 100 mi radius on our data set (see Figure 18). Nevertheless, they do not report the Average Error Distance.

[45, 42, 43] reported their achieved average error distances, with the best results provided by [43] with an AED of about 531 mi. Compared with this, with 751 km (466 mi), our approach has a much lower AED. Moreover, in parallel to our work, [115] reported that his approach achieves a median error distance of below 10 km on a worldwide data set. [94] achieved a comparable median error distance with 9 km. Nevertheless, with our approach, we achieve a median error distance of 5.05 km, thus much more precise results.

Table 12.: Comparison of our approach for estimating the user's residence with related approaches.

Approach	% within 100 mi (161 km)	AED
Cheng et al. [45]	51%	535 mi
Chandra et al. [42]	22%	1044.28 mi
Chang et al. [43]	50%	531 mi
Mahmud et al. [144]	60%	Not provided
Han et al. [94]	67%	Not provided
Our approach	79%	466 mi

The results show that with our approach, we are also able to estimate the home location of a user with high precision in a small radius around the real home location.

5.4.6 Study 4: Evaluation of Estimating the Focus of Incident-Related Tweets

In this study, we evaluated our approach for the street-level geolocalization of incident-related tweets. For the evaluation, we used Set SET_ME. The quality of our approach was estimated by comparing the ground truth geotags with our estimations. The results of this approach are shown in Table 13.

Table 13.: Results of the overall approach for estimating the focus of incident-related tweets.

RMSE	AED	MED	Recall
15.02 km	3.78 km	0.25 km	98.50%

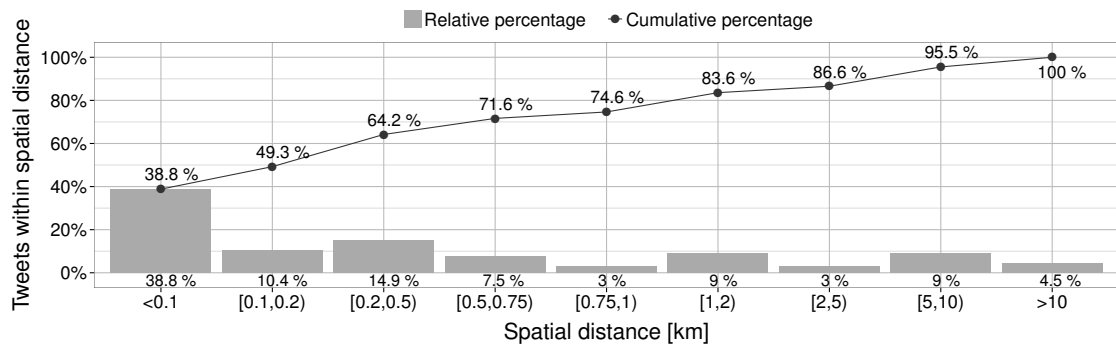


Figure 19.: Overview of the overall results for estimating the focus of incident-related tweets. The y-axis of the chart provides the percentages of the tweets within spatial distance. The x-axis shows the spatial distance in km.

The results indicate that we are able to estimate the location of an event described in a tweet with a median distance of 250 m. Unfortunately, no comparison to related work is possible as none of the related approaches [80, 222, 169] provide distance information.

Though our approach seems to be very precise, the approach can further be tuned in future work. For instance, the tweet "Accident in Bellingham - SB 5 before Lakeway Dr, right lane. Aid on scene. Slow from before Sunset Dr" has low precision as the geocoding APIs fail to geolocalize "SB 5", which is referring to the road "I-5". Future approaches should be tuned to deal with this aspect. Nevertheless, the current distance is sufficient for the street-level estimation of incident-related tweets.

5.5 Conclusion

In this chapter, we dealt with the problem of inferring the spatial dimension of a tweet [R2]. For this, we presented the first multi-indicator approach that combines vastly different spatial indicators from the user's profile and the tweet's message for estimating the point of origin for a tweet, the home location, as well as the message focus. As a result of our approach, we are able to infer spatial information for user-generated content.

In this chapter, we made the following contributions:

- We conducted an in-depth analysis of different spatial indicators that can be retrieved from tweets and identified those that are valuable for geolocalization problems using an optimization algorithm. Based on these spatial indicators, we developed several approaches for extracting location information from social media data.
- We proposed a novel approach for the geolocalization of tweets that is capable of inferring the home location of a Twitter user, the point of origin where a tweet was sent, as well as for inferring the focus of a tweet message. In contrast to other works, our method uses a large variety of indicators. Thus, it is less vulnerable to missing or incomplete data.
- We validated the accuracy of our approach on 927K tweets and showed that the approach is able to create estimations for the point of origin of a tweet for 92% of the tweets in our data set with a median of 29.66 km. Furthermore, it is able to predict the home location of a user with a median accuracy of below 5.1 km. We show that both predictions outperform the results of state-of-the-art algorithms. Also, the same approach is able to estimate the focus of incident-related tweets within a median accuracy of below 250 m.

Though our approach achieves high precision for all cases, we see further potential for optimizations. The indicators discussed may be refined (e.g., with respect to

accuracy and internal quality measures), and new indicators may be integrated in our model. For instance, Sadilek, Kautz, and Bigham [190] show promising results toward using the social network of a user for location inference. However, collecting the social network for a user is very expensive; thus, novel approaches are needed for selecting appropriate parts of the social network to collect location inference. Also, the social network may not be useful for predicting the message focus, but it could give an indication of the home location.

Furthermore, for the emergency management domain, it would be beneficial to compute an overall confidence score for the estimations to avoid false predictions. To achieving this, our introduced quality measures and the height of the final polygon may be used to calculate such a score.

In the next part, we show how we infer the thematic dimension of tweets and how we detect incidents based on user-generated content.



Part III.

Incident Detection and Clustering of Incident-Related Information



In the last part, we presented several automatic preprocessing steps. We showed how to prepare unstructured data in a way that it can be used as structured data for applying machine learning. Furthermore, we showed how we infer spatial and temporal information for each tweet.

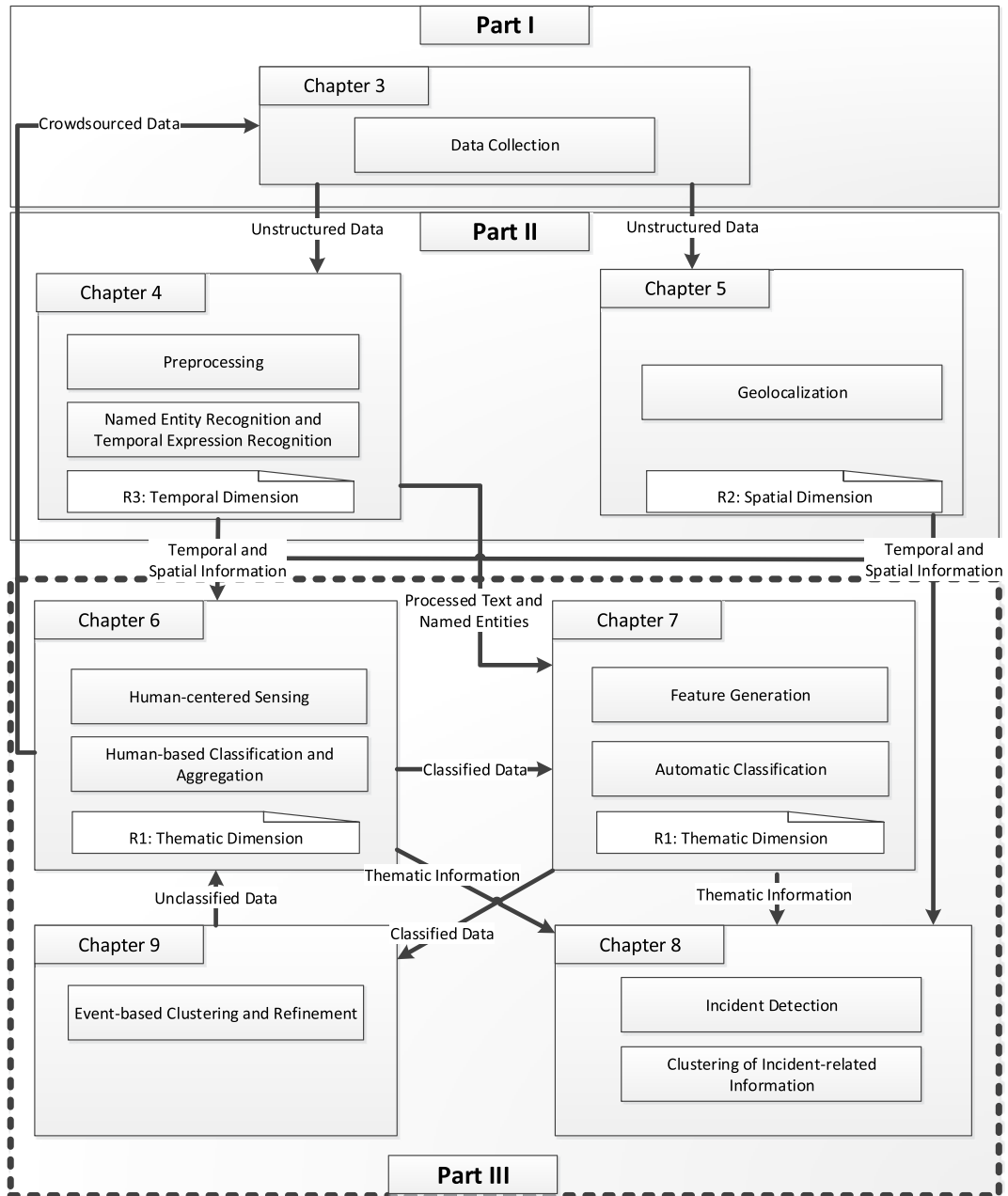


Figure 20.: Overview of the connections of Part III to the chapters in this dissertation.

The following part is divided into several chapters (see Figure 20). The goal of this part is to introduce approaches for detecting incidents based on preprocessed user-generated content and to aggregate information related to the same incident.

We deal with the requirement to infer the thematic dimension of a tweet $R[1]$. In Chapter 6, *human-based classification and aggregation* is presented, which is our approach to *manually* classify the thematic dimension of the data at hand. Furthermore, in this approach, user-generated content related to an incident is identified. As a result of this approach, valuable information can be provided as classified information base to decision makers and can be used as input for further automatic processing steps. Also in this chapter, we present human-centered sensing as a means for collecting additional information about an incident. However, crowdsourcing is costly when it comes to timely information retrieval on a large amount of information. To deal with this, Chapter 7 presents *machine-based classification*, which is applied to *automatically* infer the thematic dimension of a tweet. Also, we deal with the dynamism and regional variation of user-generated content. For this, we introduce the novel concept of semantic abstraction, which allows the creation of features that are not city-specific and support training a generalized model. As a result of machine-based classification, thematic information is obtained, which can be used to detect incidents and to cluster related information.

Based on the temporal, spatial, and thematic information, new incidents are detected. Furthermore, information about the same incidents is clustered. The respective approach is presented in the *machine-based aggregation* step (see Chapter 8).

In the last chapter of this part, we deal with the problem that social media platforms are dynamic environments. Thus, we present an approach for refining the framework according to changing conditions (see Chapter 9). This is important as the machine-based approaches need to be adapted to different incident types or different information sources. Furthermore, the refinement step helps to reduce the amount of information that needs to be processed in the *human-based classification and aggregation* step.

In summary, the contributions of this part are the following:

- We propose a general approach for applying crowdsourcing to classify and aggregate user-generated content according to the information need of the command staff in emergency management. As a result of this approach, the thematic dimension of an information item is inferred.
- We present an approach for the automatic classification of user-generated content according to the thematic dimension. For this, we propose a set of features that are most suitable for classifying the type of incident in user-generated content. Furthermore, we introduce the novel concept of semantic abstrac-

tion, which allows the creation of features that are not city-specific and support training a generalized model.

- We propose a spatio-temporal-thematic clustering approach, which is able to detect incidents based on a large amount of social media data. Furthermore, we show that we are able to detect more than 50% of real-world incidents published in an emergency management system. Furthermore, 32.14% of the detected incidents are within a 500 m radius and within a 10 min time interval of the real-world incident, allowing precise spatial and temporal localization. Compared with related approaches, the approach detects five times more incidents. Also, 77% of the event clusters created with our approach are indeed related to incidents, which shows that our approach is able to reduce noisy information.
- We evaluate the value of situational information shared in tweets posted in two North American cities. We show that a variety of individuals share information about small-scale incidents. Furthermore, we show that important situational information about affected objects, injured persons, and the location of an incident is shared.
- We present a novel event-based clustering approach for active learning that makes use of spatial, temporal, and thematic information. Evaluation results indicate that this clustering approach outperforms related strategies. Also, we show that with the presented approach, labeling costs can be reduced as less training data is needed.



6 Human-Based Classification and Aggregation of User-Generated Content

In this chapter, we deal with requirement [R1], which is the problem of inferring the thematic dimension of a tweet. With respect to this, we present how to facilitate crowdsourcing for (1) classifying incident-related information according to the thematic dimension and for (2) clustering information related to the same incident. Furthermore, we show how additional information about an incident can be collected. As a result of this processing step, classified and aggregated information can directly be provided to a decision maker or can be used in the subsequent machine-based classification step (see Figure 21).

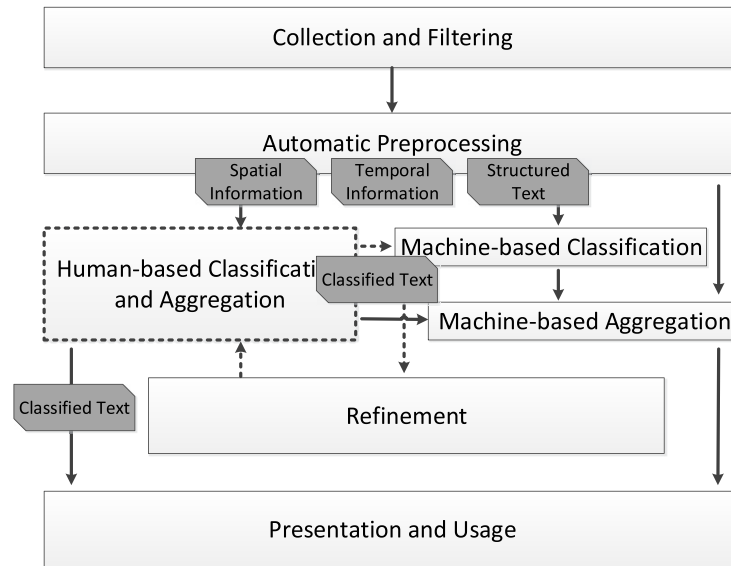


Figure 21.: The human-based classification and aggregation step in the framework.

As we show throughout this chapter, people already act as observers and share many incident-related information every day. However, nowadays, only few of the available information sources are used actively by decision makers in emergency management because of information overflow. To deal with this information overflow, crowdsourcing can be applied to classify incoming information in a way that relevant information is identified, which afterward can be provided to a decision maker. Nevertheless, it is not clear how crowdsourcing may be applied in the emergency management domain to classify incident-related information. The same applies for aggregating user-generated content related to the same incident. Thus, we need to deal with the following two questions:

-
- *How can the thematic dimension of user-generated content be manually classified?*
 - *How can user-generated content related to the same incident be manually aggregated?*

Situational information is either already present or may be collected during an incident. For this, specialized mobile applications or information in social networks may be used to gather additional information. Nevertheless, this collection needs to be related to the current information need of a decision maker. Thus, we need to answer the following question:

- *How can (additional) incident-related information according to the information need of a decision maker be manually collected?*

To deal with these questions, we present a novel approach combining crowdsourcing for the collection of incident-related information as well as for the classification and aggregation of user-generated content related to an incident.

The contributions of this chapter are the following:

- We propose a general approach for applying crowdsourcing to classify and aggregate user-generated content according to the information need of the command staff in emergency management. We show how this approach can be used to manually detect incident-related information in user-generated content and to manually aggregate information related to the same incident.
- As a result of a series of user studies, we show that this approach is indeed valuable for the command staff.
- We evaluate the quality of manual classifications of social media data and show that nonexperts and experts provide high-quality data alike.

In Section 6.1, we first distinguish different types of crowdsourcing. In Section 6.2, related approaches are presented. Our approach is shown in Section 6.3, followed by a description of a prototypical implementation in Section 6.4. In Section 6.5, an evaluation of our approach for applying crowdsourcing in emergency management is given as well as an evaluation of crowdsourcing for the annotation of incident-related information. Finally, we close with a summary and future work in Section 6.6.

Parts of this chapter appeared in [194, 196, 198]. Also, parts of the approach are patented [88].

6.1 Background

Crowdsourcing has been applied in many different areas (see for an overview [253]). In this dissertation, we focus on two application scenarios: crowdsourcing for the

collection of incident-related data and crowdsourcing for classification and aggregation of user-generated content. In the following section, we define and distinguish both types of crowdsourcing.

Participatory and Opportunistic Sensing

Over the last years, utilizing sensor-equipped mobile phones became popular for applying crowdsourcing in urban environments for collecting sensor data, which is known as "Participatory Sensing" (PS) and "Opportunistic Sensing" (OS). For both types, (sensing) *campaigns* are initiated, which are geographically and temporally constrained series of activities for capturing particular data [35, 218]. *Initiators* create such sensing campaigns, whereas people involved in sensing campaigns are called *campaign participants*. Finally, *analysts* draw conclusions upon the collected data. In our case, the initiators and analysts are employees of the command staff (e.g., the decision makers that consume information and react based on the situational information).

For this dissertation, the participatory and opportunistic approaches need to be clearly differentiated. The general idea of a participatory approach is to create sensor networks (e.g., based on mobile phones), to enable users to collect and analyze information [35]. In participatory sensing, people are in control of the sensing system, allowing them to decide when and for which campaign to collect data. In an opportunistic approach, the human is regarded as the carrier of the sensor but does not need to explicitly start the sensing activity [112]. For example, the sensor automatically collects information while the device is carried. In a participatory approach, participants have to run an application on their own, whereas in an opportunistic approach, the user only has to configure the device to let the application run. Thus, in opportunistic sensing, participants are not necessarily aware of the actual sensing activities [127], which means they do not actively react according to a certain information need. This enables more people to be information providers; however, information that is not directly relevant for a certain campaign may be collected.

Human-Centered Sensing

Nowadays, the sensing process has evolved from static sensor networks to people-centric approaches, which are based on the mobility of people [127]. The sensing activity is now conducted by a human as a virtual or human sensor. This is why this process is often referred to as people-centric sensing or human-centered sensing (HCS) [127, 218]. Compared with sensor networks, in HCS, humans are actively involved in the data collection and decision-making activities than in traditional sensor networks; humans can *sense* their environment and induce inferences, which traditional sensor networks cannot do. Thus, humans have to be treated as a versatile and unique source for information [218]. For example, they can provide real-time

information in social networks or information that is difficult to get from physical sensors such as texts. This approach of human-centered sensing has been shown as valuable for increasing the coverage over time and collecting targeted information [36, 35].

Following the initial definitions of participatory and opportunistic sensing, we define human-centered sensing as follows:

Definition III.1. *Participatory human-centered sensing is the activity of actively gathering information according to a certain information need.*

For instance, the initiator of a sensing campaign requires a certain type of information at a certain place. Thus, this type of sensing requires a directed information flow between the participant and the initiator of a campaign. By using participatory approaches, additional information about the context of the captured data can be added as metadata by the participant [35].

In opportunistic HCS, which is also known as "social sensing" [218], people are not bound to a specific campaign or application. They share information as they do it every day and ad hoc, for instance, during an accident or anything of particular interest for the individual. As a result of this, much information is created and needs to be filtered afterward according to the information need. This is the regular case when social media data is mined. Based on this, we define opportunistic human-centered sensing as follows:

Definition III.2. *Opportunistic human-centered sensing is the activity of ad hoc gathering of information without a specific information need.*

Although one major problem of sensing campaigns is the recruitment of appropriate individuals [218], the aggregation and analysis of a set of observations is still an unsolved problem [35]. Furthermore, much important information is already shared in social networks in an opportunistic manner, making novel techniques for aggregation and analysis a necessity.

Human-based Classification and Aggregation

Alongside the collection of data, crowdsourcing is also widely used for the classification and aggregation of textual information in the emergency management domain (see Section 6.2). For instance, collected information may contain spam, is outdated, or is completely irrelevant; thus, manual filtering needs to be applied. Furthermore, textual information may be *annotated* (or *labeled*) to enable training of machine learning models [145] (see Chapter 7). As people need to be explicitly tasked to create such annotations or to aggregate information according to certain criteria, we regard this as a participatory approach. In this dissertation, we refer to

this approach as *human-based classification*. Furthermore, classifying data according to the same thematic dimension enables the aggregation of related information. Though the crowdsourcing task is different, the previously presented definitions for crowdsourcing campaigns apply likewise.

Summary

In this section, we defined two approaches for crowdsourced collection of data, namely, participatory human-centered sensing and opportunistic human-centered sensing. We showed that both differ in the way participants are tasked. Furthermore, we presented human-based classification as a means for applying crowdsourcing for manual analysis of collected data.

6.2 Related Work

Various approaches exist for applications that make use of user-generated content and that apply different means of crowdsourcing in the emergency management domain. Those approaches differ with respect to (1) information sources and (2) processing methods.

On the one hand, related applications usually comprise many different *information sources*:

- Static web content such as text snippets collected from websites.
- Opportunistic human-centered sensing such as Twitter messages or pictures on Flickr.
- (Open) Government data (e.g., governmental information about incidents³³).
- LOD as a source of interlinked information.
- Traditional sensors such as water level or earthquake sensors.
- Participatory human-centered sensing (e.g., based on specialized mobile applications).

This information needs to be further *processed* in order to get meaningful insights from the raw data. This includes approaches that make use of human-based classification and approaches making use of machine learning for filtering. Information sources are automatically or semiautomatically enriched with further knowledge. As these approaches are also related to the works presented in Chapter 7, we give an overview of processing algorithms and evaluation results.

On the other hand, processing methods are differentiated into the following areas:

³³ See <http://www.data.gov/> [Accessed: 11.02.2014] for an example.

-
-
- Natural Language Processing methods are often used for extracting relevant topics and/or information snippets from text.
 - Semantic annotation is used for tagging and linking information items with further metadata and external knowledge such as LOD sources.
 - Machine learning can be employed for further processing the data (e.g., for clustering incident-related information).
 - Simple information filtering is applied based on metadata (e.g., using temporal or spatial information).
 - Human-based classification is applied if automatic means are not applicable (e.g., by asking humans to categorize pictures).

Table 14 summarizes the approaches discussed in this section and shows which information sources and processing methods they comprise.

In [109], Sheth et al. presented an approach using Twitter for sense making (e.g., for the identification of events in tweets). Twitris³⁴ follows an opportunistic human-centered sensing approach. The application extracts information about real-world events provided by citizens and presents related information to these events. Incoming text is analyzed based on spatial, temporal, and thematic dimensions. These dimensions are used to cluster tweets according to events based on text similarity using TF-IDF scores. The differences in similarity are then used to weight the relevance of a tweet for an event. Though the platform follows a similar approach as this dissertation, no evaluation results are available. Furthermore, the authors admitted that the spatio-temporal-thematic analysis is done with a week of lag.

Heim and Thom [96] proposed SemSor for supporting the situational assessment for emergency management based on opportunistic human-centered sensing and automatic social media analysis. The goal of this project was to automatically identify background information for incident-related social media data. The approach is based on the constant crawling of social media sources like Twitter, Flickr or YouTube. The textual entries of the social media items are annotated with links to entities in LOD. The automatic analysis is based on spreading activation [51], which is applied on LOD to identify related entries that might be useful. Further, no evaluation was provided for this application.

Likewise, [150] presented the VisInfluence application for visualizing sentiments related to topics such as "earthquake". The approach relies on extracting named entities and sentiments in Twitter messages using the Alchemy API³⁵. In contrast to SemSor, VisInfluence focuses on the intelligent visualization of topics; thus, no further processing is applied.

³⁴ Twitris is available on <http://twitris.knoesis.org> [Accessed: 22.05.2013]

³⁵ <http://www.alchemyapi.com/> [Accessed: 22.05.2013]

Table 14.: Overview of approaches that apply crowdsourcing for processing incident-related information.

Approach	Information Sources					Processing Methods					
	Static Web Content	Opportunistic HCS	Government Data	Linked Open Data	Sensors	Participatory HCS	NLP	Semantic Annotation	Machine Learning	Information Filtering	Human-based Classification
Twitris [109]		X		X		X	X	X	X		
SemSor [96]		X		X				X	X	X	
VisInfluence [150]		X		X				X			
Disaster 2.0 [178]						X					X
Linked Sensor Middleware [129]				X	X			X		X	
Twitcident Abel et al. [1]		X	X	X			X	X		X	
SaferCity [25]		X	X				X		X		
Repopulation Indicators for New Orleans [87]			X								
IBISEYE [97]					X						
EDIS [159]			X		X					X	X
Live Earthquake [47]					X						
Healthmap [34]	X									X	
MediSys [114]	X	X	X				X			X	
LA Fire Tweets [9], SwineFluTweets [10], Iran Protest Tweets [11]		X									
WikiCrimes [233]						X				X	
LA Fires [230], Bushfire Incidents [67]	X		X							X	
PakReport [182]						X					
Ushahidi [165]		X				X				X	X
FindShelter [77]	X					X				X	
Our approach		X		X		X	X	X	X	X	X

The Disaster 2.0 system followed the idea of using a participatory human-centered sensing approach and applying human-based classification to manage and filter information about natural disasters [178]. To that end, the main entities like events (fire, flood), allocated resources (policemen, firemen), and damages (victims) are explicitly modeled in the system with unique identifiers. Information about those entities are then obtained by explicitly asking citizens (e.g., using a specialized mobile application for PHCS). Human-based classification is also used to refine existing information about an incident such as the correct position of resources. No further processing is conducted of the information at hand.

The Linked Sensor Middleware (LSM) platform was presented as an example that combines static information sources such as sensors as well as user-generated content provided in the Semantic Web based on Linked Data principles [129]. The purpose of the middleware was to make sensor streams usable by integrating them with existing information. The whole concept is based on the idea of Linked Stream Data [206] to provide a way to publish heterogeneous sensor data as Linked Data. Based on a remote SPARQL endpoint, automatic filtering of the information at hand is enabled. The Live Linked Open Sensor Database project [130] is based on the Linked Sensor Middleware, with two application scenarios for emergency management mentioned: by using sensor data, additional information from these sources might be taken into account for decision making (e.g., the wind force or temperature in case of a fire). Furthermore, the combination of user-generated content and sensor data can be used to prevent the spread of a disease.

Abel et al. [1] presented Twitcident³⁶ as an application that allows the filtering, searching, and analyzing of social media data about incidents. For incident detection, information about incidents published in the P2000 network³⁷ is collected, which is used by the public emergency services in the Netherlands. Therefore, the type of incident, its location, and its temporal attributes can be retrieved. The Twitcident framework transforms the information into an initial query for collecting potentially relevant Twitter messages (tweets) using the Twitter Streaming API³⁸. For filtering relevant information, several steps are conducted [2]: First, based on the incident information, an incident profile is created, which describes the attributes of the incidents. This integrates related concepts like locations or persons and weights their importance for the incident. As a second step, tweets are collected based on the incident profile. Then NER is applied to detect entities like persons and locations in tweets. Fourth, a classification component classifies each tweet into different types, using categories such as different damages and risks. For further filtering, keyword-based and semantic filtering is applied to automatically identify relevant tweets for

³⁶ <http://wis.ewi.tudelft.nl/twitcident/> [Accessed: 22.05.2013]

³⁷ [http://en.wikipedia.org/wiki/P2000_\(network\)](http://en.wikipedia.org/wiki/P2000_(network)) [Accessed: 22.05.2013]

³⁸ <https://dev.twitter.com/docs/streaming-apis> [Accessed: 22.05.2013]

an incident. The framework has not been evaluated with respect to its capability of incident classification and aggregation.

SaferCity [25] is also based on opportunistic human-centered sensing using Twitter data. All geotagged tweets are clustered before assigning a type to each cluster. For this, incident-related keywords as well as the message content are used for applying a machine learning approach. Furthermore, sentiment analysis is applied to assign a sentiment score to each cluster. In addition to this, they proposed to collect governmental data about incidents. Nevertheless, it remains unclear how this information is used in the overall approach. Though the approach was applied on tweets collected in New York City, no evaluation is provided.

The Ushahidi platform is an open-source crisis management application that was developed in 2008 [165]. The application is the most popular example of how opportunistic and participatory human-centered sensing are combined. Ushahidi follows an opportunistic HCS approach for incident reporting and became known worldwide of the Haitian earthquake, when people were asked to translate Creole text messages in a participatory manner. The platform has been used for several campaigns over the last years (e.g., for tracking events around the Gaza Strip, as shown in Figure 22).

With the increasing amount of data on the platform, Ushahidi, Inc. developed the SwiftRiver toolset as a complimentary product [86]. The SwiftRiver toolset provides several APIs for filtering and structuring data from multiple information sources. Social media, blogs, or mobile applications can be used as an information source. The fetched data can be processed by different APIs. For example, the location of the information is identified. Content is analyzed lexically, and duplicates are removed. As a result, contextualized messages are created, which can be processed by asking questions like "Where is the person at?" or "What has happened?" based on the predefined tags provided in a taxonomy. Though the application has shown its value in real-world examples many times, no evaluation of the SwiftRiver tools is provided.

As another approach, Fritz et al. [77] proposed FindShelter. The application is based on crawling web pages for important information like emergency accommodations as this source of information cannot easily be aggregated by decision makers. For filtering, relevant information is crawled based on manually defined web pages. Then regular expressions are used to extract the relevant information (e.g., for phone numbers and addresses). Finally, data from different websites are merged and published as RDF and as an OGC-standard web feature service and can be displayed on a map view.

Besides applications comprising comprehensive filtering, other approaches focused on combining and visualizing information provided by human sensors. The use of recently updated information sources like publicly available information, govern-

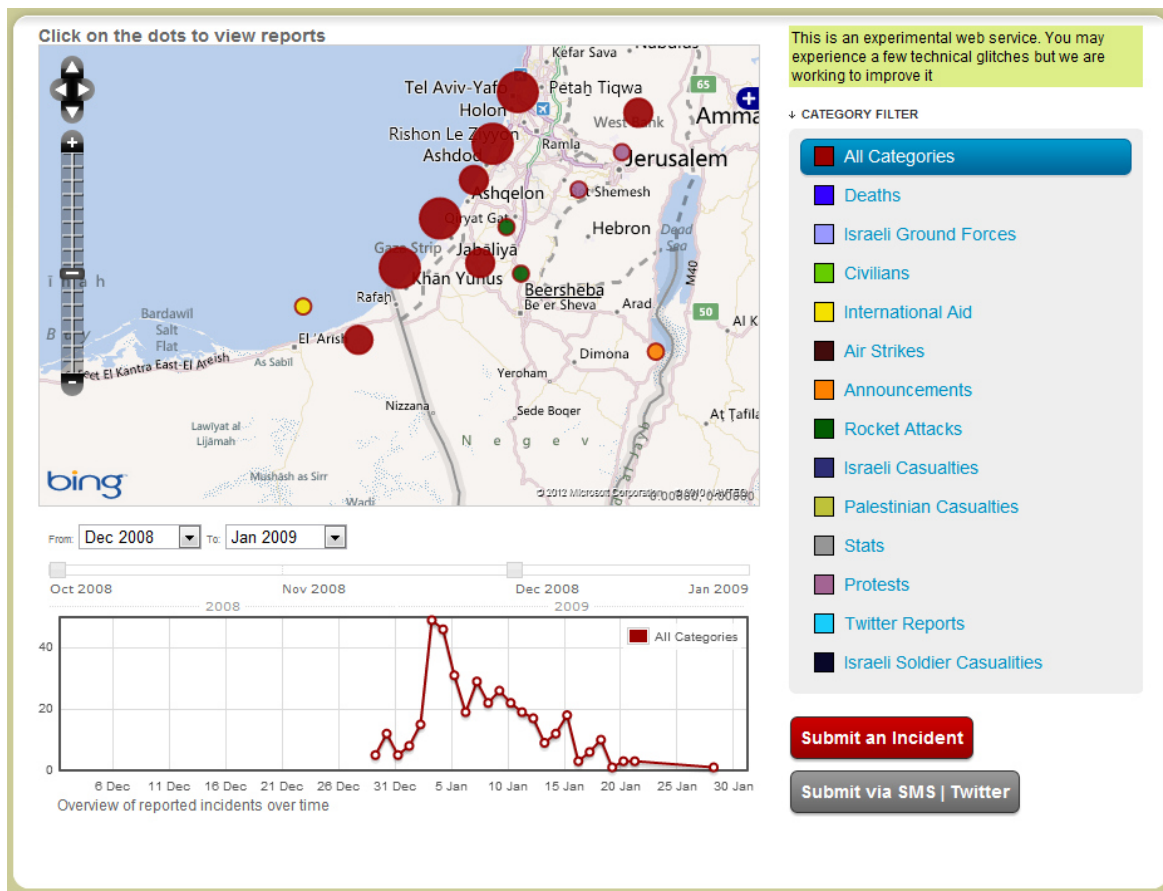


Figure 22.: Screenshot of the Ushahidi platform for tracking events in the Gaza Strip.

mental data, or crowdsourced information enables the view on spatial and temporally changing events. Thus, crisis map mashups can be used to report, assist, and manage emergencies. Zang et al. [254] describe map mashups as a common form of data mashups because they are "the most visual and adaptable of the mashup options". Several analyses on crisis map mashups have been made (e.g., by [142]).

Generally, crisis map mashups can be roughly categorized into mashups that show information from only one source (e.g., Twitter messages on a given topic), and mashups that aggregate information from different sources. For the first category, examples such as [87, 97, 9, 10, 11, 40] focused on visualizing different information on a map. For instance, weather alerts, distribution of flues, or reports about incidents were visualized. In the second category, applications such as [159, 230, 67, 182, 34, 233, 114] employed some intelligent processing for aggregation (e.g., topic and entity detection for relating pieces of information from different sources). However, the focus of these applications is more on the visualization aspect and not the processing methods.

Summary

In this section, we showed that existing approaches make use of a variety of information sources for gathering information for emergency management. In particular, the combination of opportunistic and participatory human-centered sensing is widely used by combining information provided in social networks with information provided by specialized mobile applications. Furthermore, integrating static web content, LOD, or traditional sensors in some cases augment the initial information base.

The collected information is further processed in most cases. The metadata is used frequently to filter the information base. NLP, machine learning, and semantic annotations are also combined for processing. Human-based classification is only applied in a few cases. Though different processing methods are used, the combination of applying human-based classification and automatic means such as machine learning is rare. In contrast to existing approaches, our framework relies on both: crowdsourcing as well as on different mechanisms for automatic processing of user-generated content. Furthermore, though most examples have been implemented successfully, no approach was evaluated with emergency management personnel.

6.3 Approach

In the following section, we present our approach for human-based classification and aggregation of user-generated content. As a starting point, either in participatory or opportunistic HCS, decision makers act as initiators of sensing campaigns; thus, they must be allowed to express their information need. To facilitate this, we follow an approach that is commonly used in question and answer communities. The idea of these communities emerged from simple forums that allow people that have an issue to ask for advice by posting a question to the community. For each question, the community tries to post possible answers to solve the question. This approach has been shown quite successful and still can be found in communities like Ask.com³⁹ or Stack Overflow⁴⁰.

For setting up an HCS campaign, the same mechanisms can easily be adapted. If a decision maker needs to take a decision, he can ask questions to a community ("the crowd"). The community can provide helpful information like texts, pictures, or videos that might help to answer a question. In our case, this question is related to identify all information related to the same incident. Nevertheless, question asking allows a much more precise aggregation of information related to the same incident. In the example process shown in Figure 23, we differentiate the point in time when the questioning takes place as this is depending on the type of HCS campaign.

³⁹ ask.com

⁴⁰ stackoverflow.com

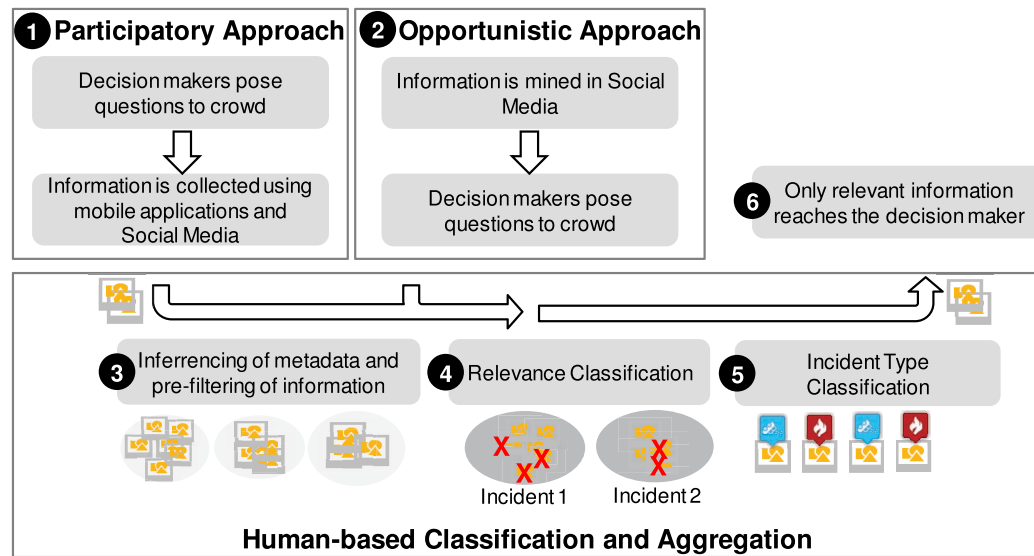


Figure 23.: Example process for human-based classification.

In a participatory approach (Figure 23, ①+②), asking a new question initiates two campaigns: a campaign for collecting incident information and a campaign for classifying the incoming or present information. As a result, campaign participants collect and report information that might be helpful for the sensing campaign using, for example, mobile applications or social media platforms. As this information is directly related to the incident, it is called an *incident report*. Depending on the mobile application, metadata such as GPS coordinates or specific incident types might be added by the user itself. Afterward, based on the incoming information a crowd can classify information in the filtering campaign according to a certain information need. In this approach, the crowd for sensing and the crowd for classification do not necessarily consist of the same people.

In an opportunistic approach (Figure 23, ①+②), information was already created without a specific campaign, for instance, in social media platforms. In this case, the question directly initiates a human-based classification campaign and is directed toward a crowd for classifying the information according to the question. In the opportunistic approach, regular information items need to be separated from valuable incident reports. In this approach, only metadata that was already collected by the social platform can be used, which makes further processing steps a necessity.

As a next step, the prefiltering of the information at hand takes place (Figure 23, ③). Since metadata might not be present for the collected information, it needs to be inferred automatically. For this, the approaches presented in Chapter 4 can be used to infer spatial and temporal metadata. Furthermore, as in both cases the incoming

information remains mostly unfiltered, a prefiltering using the inferred metadata can be conducted.

Based on this prefiltered information, the participants of the human-based classification campaign act as reviewers to filter the incoming information according to their relevance (Figure 23, ④). In this case, if information is classified as relevant, this means that it is helpful for answering a question (i.e., incident reports are separated from noise). Additionally, the crowd can classify the information items according to incident types (⑤) (i.e., assign an incident type such as "fire incident") to an information item. For instance, an image might be related to a car accident while some text is related to a fire. Finally, the filtered information can be used by a decision maker (⑥).

In the following, we present how this approach is adapted for our framework (see Figure 24).

Step 1: Initiating a human-based classification campaign by asking a question

As mentioned before, our goals are (1) to manually classify the thematic dimension of user-generated content and (2) to manually aggregate user-generated content related to the same incident. This finally reduces the incoming flood of information (Figure 24, I1) to a set of information that is directly relevant to suffice the information need of a decision maker (e.g., the command staff). Based on the idea presented before, we choose the approach of having the command staff asking questions in order to articulate a particular information need (Figure 24, D2)⁴¹. The incoming information, either from a participatory or opportunistic approach, is classified according to its usefulness to answer such a question. We call this approach *Question Guided Relevance Rating*.

During interviews with command staff members, the following example questions were considered to express realistic information needs according to specific incidents:

- "Does the smoke contain poisonous chemicals?"
- "Can we cross the bridge with a 12-ton fire truck?"
- "Is the road accessible?"

Posing these questions is not trivial: It requires clearly expressing the information need and transforming it into a short and concise sentence. First of all, questions must be comprehensive for citizens as well as domain experts. Both user groups must be able to infer the part of the information from the question that is appropriate with respect to their skills. Second, a question must not contain any ambiguity that could distort feedback from the crowd. To deal with this, ambiguity could

⁴¹ In this case, the participatory HBS approach is shown.

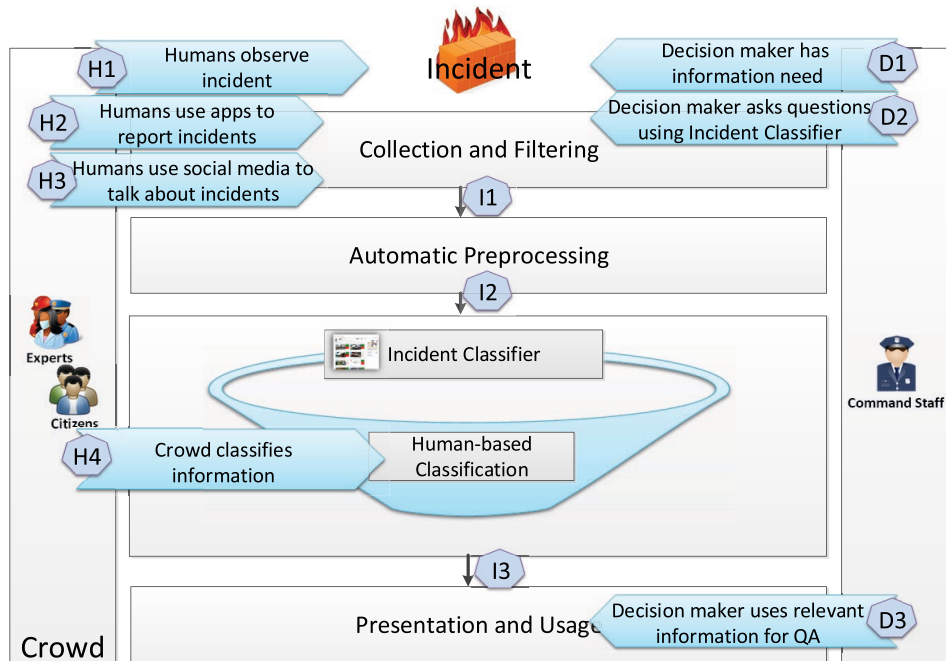


Figure 24.: Participatory human-centered sensing and human-based classification integrated into the overall framework.

be detected automatically by measuring the inter-user agreement on classifications. Finally, a question must be precise. This implies not only an avoidance of redundancy but also a hard instead of fuzzy edge between the sets of relevant and nonrelevant information.

Furthermore, to correlate a question with a certain incident (Figure 24), the decision maker has to define the spatial extent of the area a question applies to as well as the temporal extent to define how long a question is valid. For spatial boxing, the decision maker defines a spatial bounding box for which the question applies. Thus, questions are only valid in a certain area, ensuring that only information that is helpful is collected. For temporal boxing, the decision maker defines a time interval for which the question is valid or applies to. For instance, the question "Is the road accessible after the crash at 9:00 p.m.?" might be spatially bounded to a certain intersection.

Step 2: Preprocessing the incoming information

As a first step in the processing pipeline, the automatic processing presented in Chapter 4 takes place. Thus, by using spatial and temporal information extracted from user-generated content as well as the predefined properties of a question, completely irrelevant information is filtered out while other information is regarded as suitable

to be manually rated as it is likely related to the event. Hence, in this prefiltering step, I1 can be filtered either using the inferred metadata or the metadata that was set by the user. As a result, a prefiltered information base I2 (Figure 24) is created.

Step 3: Human-based Classification

Based on the prefiltered information base, human-based classification is applied to analyze the information and to classify it according to the relevancy for a specific question (Figure 24, H4). As shown in Figure 23, we differentiate between two different types of classifications:

- **Relevance Classification:** A binary decision whether information is relevant for a question. Based on this approach, information related to the same event can be identified.
- **Incident Type Classification:** Classifying of user-generated content according to predefined incident types. This approach is much more restrictive but provides the thematic dimension for the information at hand.

For classifying information items according to their relevance for a question, observers or volunteers are not tasked to answer those questions directly; thus, it is much easier for them to decide whether some pieces of information could support finding the correct answer to the question. For example, when dealing with the question "Does the smoke contain poisonous chemicals?", a nonchemist may have difficulties in deciding whether some smoke is poisonous, but they may select pictures or pieces of text concerned with the color and smell of that smoke. Hence, the crowd must not necessarily be trained rescue personnel in order to produce useful classifications. When it comes to classifying information according to the thematic dimension (i.e., to pre-defined incident types), it was shown that the labeling quality might be dependent on the domain knowledge of the annotators [17], [255]; thus, this task is much more difficult. We deal with both problems in the evaluation in Section 6.5.

After these processing steps, the information base I2 is turned into I3, containing only such user-generated content that is potentially related to the same event, thus relevant for the command staff.

Step 4: Using the classified information

The result of the previous steps is a data set (Figure 24, I3) containing the information items resulting from the human-based classification and aggregation step. The information items are now related to a specific incident or have been assigned a thematic dimension (i.e., an incident type). These can now be used by the decision maker, who can consume all available information according to his information need. He is able to review the provided information according to his questions and

accepts a question as answered. With the prefiltered, aggregated, and classified information base, the decision maker is now able to take better informed decisions because he gained an improved understanding of the situation at hand.

6.4 Prototypical Implementation

In Figure 25, the user interface for the implemented prototypes is shown⁴². ① depicts the user interface for posing new questions including the spatial and temporal boxing (i.e., the correlation to a specific incident). It is shown how a user can enter a question and select the spatial location as well as the temporal interval. Furthermore, he is able to see other questions that were recently asked. Also, he can share the newly posted question to several social networks to increase the number of responses.

Figure 25, ② shows the "Incident Reporter" mobile application for incident reporting, which allows the submission of images, audio, and textual descriptions. Furthermore, a submitted report contains metadata like spatial and temporal data. The spatial metadata can further be enhanced by allowing the reporter to set the point of origin where the reporter currently is and the point of interest that the image represents. In this case, for analysts of information, it is easier to differentiate where the incident is truly located and to infer the angle from which the picture was taken. The reporter information is then sent to the "Incident Classifier" application.

Figure 25, ③ shows the "Incident Classifier" application, which is based on the question guided relevance rating. Using a web browser, people can access the Incident Classifier from anywhere. They can select a specific incident and one of the questions posed by the command staff. The classifier provides the prefiltered information so that each information object can be rated with respect to its usefulness for answering a question. Figure 25, ④ shows examples of images and texts that are to be rated with respect to the question shown in the header of the screen.

6.5 Evaluation

In the following section, we present the results of two evaluations: The first evaluation (see Section 6.5.1) aimed at evaluating our approaches for participatory human-centered sensing and human-based classification. In the second evaluation (see Section 6.5.2), we studied the quality of human-based classifications of user-generated content.

⁴² The shown prototypes were developed in the context of the research project InfoStrom funded by the BMBF (13N10712).

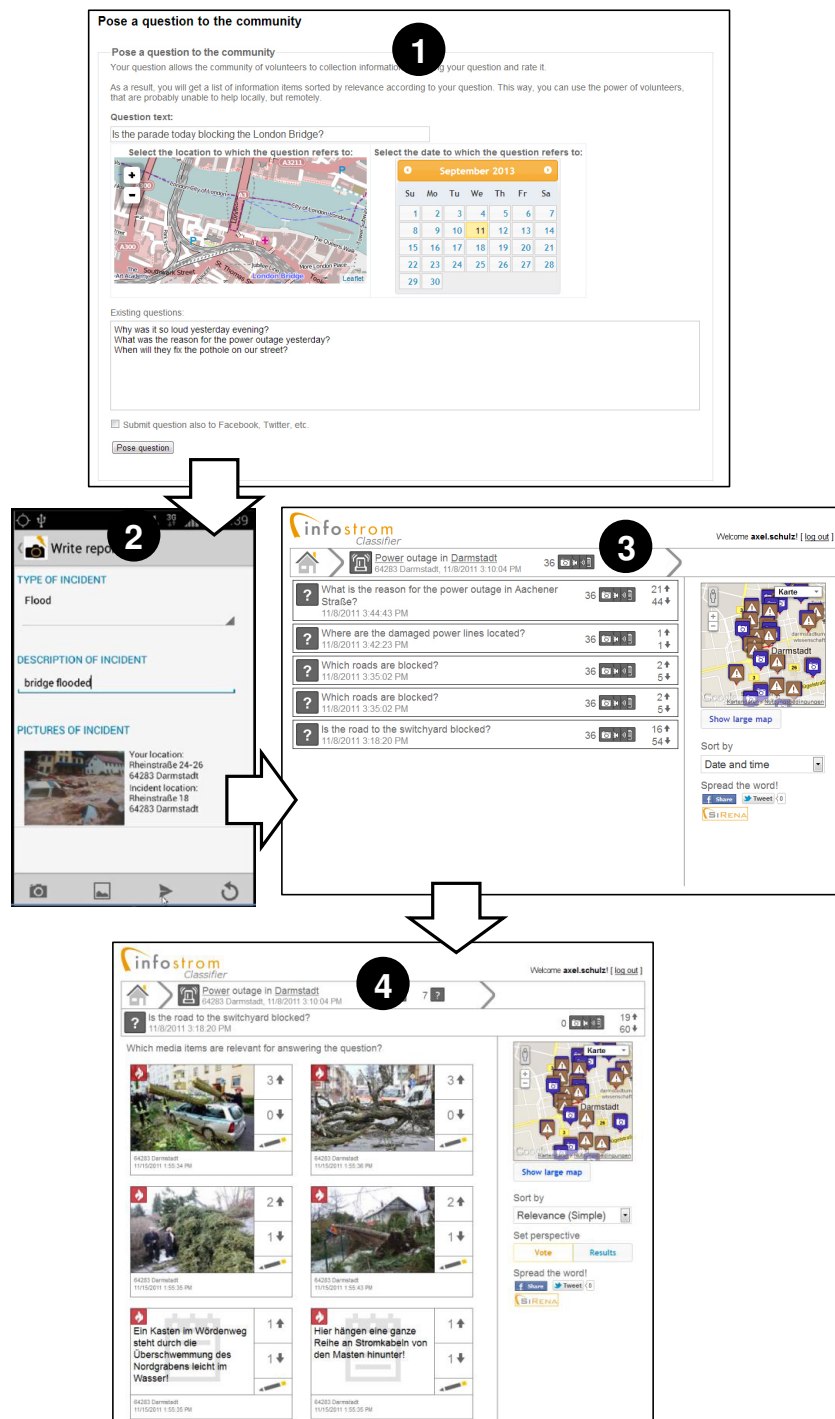


Figure 25.: Screenshot of a prototype leveraging participatory human-centered sensing and human-based classification for emergency management.

6.5.1 Study 1: Evaluation of Human-Centered Sensing and Human-Based Classification in Emergency Management

For analyzing the validity of our approach, in 2012, we conducted a series of user studies with experts from the emergency management domain. First, we wanted to evaluate if incident reports are valuable for decision makers in emergency management. As a second goal, we wanted to understand if the approach presented in this chapter is capable of reducing the incoming flood of information and, third, if it helps to reduce it to a set of valuable information for decision making.

This evaluation was conducted with the prototypes presented in Section 6.4. We did not include any other related work as this was not available in our project context.

6.5.1.1 Evaluation Design, Method, and Procedure

The studies were designed to take place for two hours. We conducted three tasks to demonstrate different aspects of our prototypes:

- Task 1. Creating User-Generated Incident Reports: The participants had to create incident reports using the Incident Reporter application. For this, the participants were introduced into a scenario ("flooding after thunderstorm") and were provided with example pictures in the form of printouts. The users had to take a picture, set a text describing the situation at hand, and provide metadata such as the incident type as well as the GPS coordinates of the incident location.
- Task 2. Rating of User-Generated Incident Reports: Based on the reports that were sent in Task 1, the users had to rate the information according to pre-defined questions (e.g., "Which roads are blocked because of flooding?"). For this, the users were introduced to the Incident Classifier and how to classify reports. Afterward, the reports were classified by the users with respect to their relevance for the questions.
- Task 3. Understanding the Situational Picture: The users should act in the role of a decision maker to gather an understanding of the situation at hand using the information provided after Task 1 and Task 2. For this, the users had to use the Incident Classifier to evaluate if the classifications resulting from Task 2 are useful for satisfying their information need.

At the beginning of the workshop, the participants were introduced to the study and the background of the scenario. For this, handouts with descriptions of all tasks and how to proceed were distributed. The participants brought their own Android

devices, and we installed the corresponding mobile applications to their system. The rest of the participants were provided with preinstalled Android devices.

Prior conducting the tasks, demographic data about the users' characteristics such as gender, age, and expertise have been collected. After each task, the users were handed out a questionnaire for evaluating their user experience as well as to further feedback in a semi-structured interview. For evaluating the first two tasks, a (1) User Experience Questionnaire (UEQ) [128] for analyzing six factors (i.e., attractiveness, perspicuity, efficiency, dependability, stimulation, and novelty) as well as (2) a couple of fill-in-the-blank text questions were part of the structure. For the third task, only fill-in-the-blank text questions had to be answered as the application was the same as in Task 2; thus, the UEQ did not need to be evaluated again. Furthermore, we took notes during the interview for the main statements.

We conducted the three studies with a total amount of 23 participants who were invited via e-mail. The participants were professional members of the fire departments of Siegen and Bergheim (8 participants), both towns in North Rhine-Westphalia, the corresponding police departments in both towns (9 participants), or work for the utility provider RWE (6 participants). The participants had 2 to 40 years of experience in their domain (see Table 15).

Table 15.: Years of work experience of user study participants.

Yrs Exp.	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40
Participants	5	4	3	1	0	1	8	1

6.5.1.2 Results

Hereinafter, we present the result of our user study. For analyzing the results, we followed the approach of [82] to separate the quantitative results to intervals: a "neutral" interval that contains values between -0.8 and 0.8, "very good" and "very poor" with ratings higher than 2 and smaller than -2, and a "good" and "poor" interval with the in-between values. Also, the statements were clustered manually to identify those statements that were most frequently given by the participants.

As an overall result of the evaluation, we found that the initial motivation of this dissertation (i.e., for some incidents not enough situational information is present) was confirmed by the participants. Also, reducing information overflow based on human-based classification was positively perceived. Furthermore, the general impression of the prototypes was that the "presented information is well structured" and the "applications are useful" in the daily work of an emergency response staff.

Task 1. Creating User-Generated Incident Reports

The participants stated that a mobile application for incident reporting would be useful for gathering information during severe storms (5), large-scale incidents (4), traffic incidents (4), power outages (3), or fires (2). This shows that for large-scale incidents as well as for small-scale incidents user-generated content is perceived as useful alike. Also, the participants thought that sending user-generated incident reports is indeed valuable for improving the situational picture (yes (20), no answer (3)).

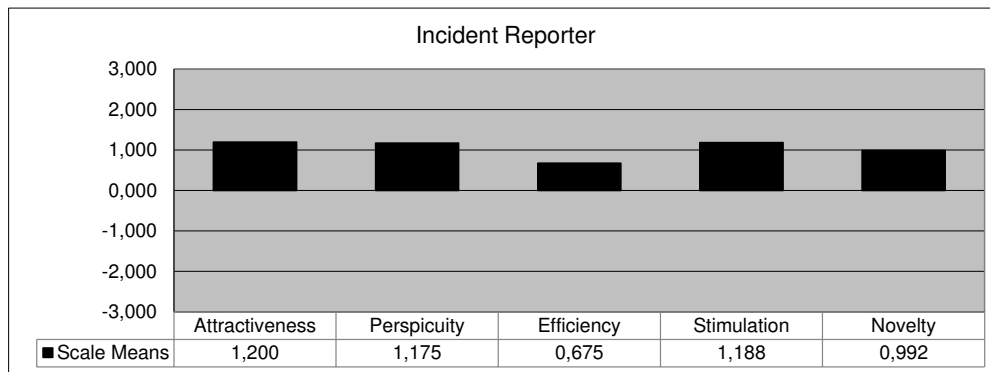


Figure 26.: The results of the user experience questionnaire for Task 1.

Table 26 shows the quantitative data that could be achieved for the dimensions UEQ for the reporter application. The results show that good results could be achieved for attractiveness, perspicuity, stimulation, and novelty. However, efficiency was rated as neutral, which, as we have shown, is likely a result of the performance of transferring information to the classifier application.

The participants pointed out the importance of adding metadata such as GPS coordinates and predefined incident types to the information at hand. Also, the general application design was perceived as fast and responsive, and the general structure of the application was appreciated. The participants disliked the slow transfer rate, which is a result of sending large images over the GPRS network. This is the main factor for the neutral efficiency rating. Furthermore, participants wanted to access already-reported information on their mobile device, which was not available as a feature.

Task 2. Human-Based Classification of User-Generated Incident Reports

The participants stated that an application for human-based classification of user-generated incident reports would be useful for filtering information regarding large-scale incidents (5), severe storms (3), power outages (1), sudden crowd gatherings (1), and for planning response measures (1). Also, the participants think that *fil-*

tering user-generated content is important for improving the situational picture (yes (17), no answer (6)).

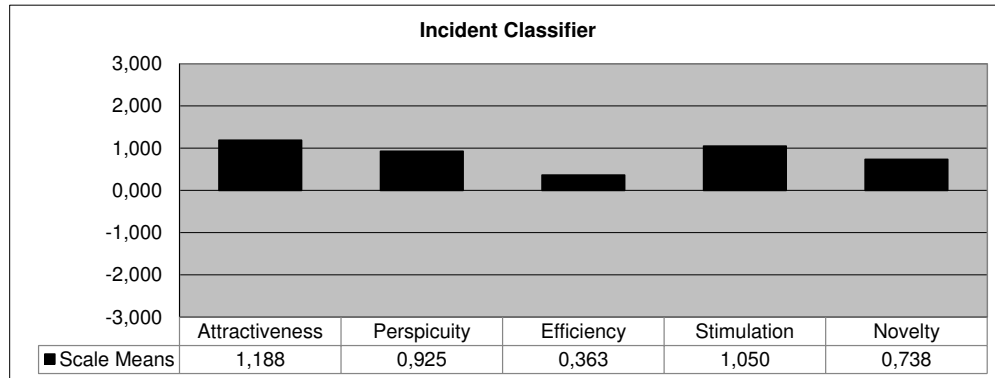


Figure 27.: The results of the user experience questionnaire for Task 2.

Table 27 shows the quantitative data that could be achieved for the dimensions of the UEQ for the classifier application. Good results could be achieved for attractiveness, perspicuity, and stimulation. However, efficiency and novelty were rated as neutral. Also in this case, the results on efficiency are neutral because of the performance aspects of our prototypes.

The participants liked our approach of aggregating user-generated content related to an incident according to their relevance for answering questions. Also, they appreciated the ability to sort items according to their relevance. Furthermore, the visualization of different incident types was perceived as useful as the different organizational responsibilities could be addressed in this manner. The spatial geolocalization of information items on a map was mentioned as another important feature.

The participants disliked the way images and their corresponding textual annotations are presented. For the participants, both should be taken as one item, whereas we think that the differentiation between an image and its text is important as the text may not be valuable for answering a question, but the image may show important aspects. Furthermore, the participants criticized minor usability issues.

Several missing features were identified by the participants; for instance, users wanted to add additional comments to their classifications to explain their decision. Furthermore, participants would like to decide between multiple levels of classifications and did not want to be restricted to a binary classification. Also in this case, the ability to get in contact with a campaign participant was mentioned as an important but missing feature. Furthermore, a filtering based on spatial and temporal aspects and additional prefiltering were wanted. Finally, it was mentioned that such a system should be able to run locally as during large-scale incidents, power outages may limit Internet access.

Task 3. Understanding the Situational Picture

From the point of view of a decision maker, the human-based classification step and the spatial boxing were mentioned as the most valuable aspects of our application. Also, the combination of textual and image information was perceived as an important element as the situational picture would be much better in this manner. However, the participants had concerns regarding the performance of such a system during times of crisis. For example, with the shown prototype, the performance dropped significantly when multiple users accessed the system.

It was also noted that the way of sorting information items according to their relevance for a question could be confusing for a decision maker as items showing incidents that are not relevant for a question would be filtered out. Thus, a combination of manual question asking, inferring, and showing potential incidents is needed. For further developments, concepts for the reputation of individual users should be integrated. Also, the different roles of campaign participants (e.g., as part of the crowd or the rescue squads) should be visualized.

Finally, human-based classification was perceived as correct and suitable for decision making (yes (14), no (1), no answer (8)). Furthermore, our approach was perceived as applicable for daily use (yes (15), no answer (7), no (1)).

6.5.1.3 Study 1: Summary

In the evaluation, we analyzed several aspects of applying participatory human-centered sensing and human-based classification in emergency management based on prototypical implementations. Throughout this evaluation, we showed that user-generated content is perceived as a valuable source of information for improving the situational picture of emergency staff. Also, additional information such as a specific location or incident type is important information alike. We also found that filtering is an important aspect to make user-generated content usable for decision makers. Our approach for human-based classification and aggregation based on the *question guided relevance rating* was perceived as one appropriate means for this. The evaluation with emergency personnel gives us an indication that our approach, which combines spatio-temporal filtering with human-based classification, is indeed valuable for real-world application.

6.5.2 Study 2: Qualitative Evaluation of Human-based Classifications

In the first evaluation we showed how to apply participatory human-centered sensing for collecting user-generated content in emergency management as well as how the classification of this data may be applied. In this evaluation, we focus on evaluating the quality of human-based classification as the provided thematic dimension

directly influences automatic approaches that make use of user-generated annotations. Compared with most approaches that assume perfect annotation quality, we wanted to quantify error rates for human-based classifications. First, this gives us an indication that human-based classifications are an appropriate means for annotating data sets for emergency management, where high-quality data is a necessity. Second, we wanted to make use of these error rates for further evaluations in Chapter 9 as these are fundamental for developing novel machine learning approaches.

6.5.2.1 Definitions

When it comes to annotation (or labeling) of user-generated content such as tweets, two problems have to be differentiated: *lack of attention* and *multi-annotation* problems. For the first category, the annotation quality does not substantially depend on the background knowledge or expertise of the current annotator. Problems of this type mainly occur due to slips [163] on the annotator’s side. For the second category, problems originate from tweets that are difficult to annotate. For instance, tweets are related to multiple incident types; thus, it could be classified with multiple annotations. For example, the tweet

Listing 6.1: Ambiguous example tweet.

THIS CAR HIT THE FIRE HYDRANT AND CAUGHT FIRE....SOMEONE HOLIDAY
ALTERED

could be labeled with multiple incident types (“fire incident” and “car incident”). When annotators are forced to decide for one single label from a set of labels of which multiple could be assigned, the lack of domain knowledge could lead to an increased error rate. Thus, we conducted our study with different user groups: *domain experts* and *nonexperts* with no or limited domain knowledge.

As both problems result in different error rates, two types of error rate have to be differentiated [255]: *random error* and *systematic error*. The random error results from the annotator’s carelessness, which is a result of lack of attention. For example, a wrong label is occasionally assigned. The random error is regarded as i.i.d. noise on each label; thus, we assume a fixed probability $RE \in [0, 1]$. The systematic error results from labeling samples that are difficult to annotate, which are multi-annotation problems. In this case, the label noise is correlated. The systematic error is a probability $SE \in [0, 1]$. We further assume that either a random or a systematic error occurs for each tweet; thus, we have a set with ambiguous tweets and tweets for which a random error occurs. In the following study, we also report the overall error (OE), which is the labeling error for all tweets.

6.5.2.2 Hypotheses

Following related work, we assume a difference in labeling quality as the quality might be dependent on the domain knowledge of the annotators in crowdsourcing environments [17, 255]. This results in three hypotheses:

- $H1$: The means (μ) of the *overall error* are different across both user groups:
 $H1_0 : \mu_{OE,CU} = \mu_{OE,EX}, H1_A : \mu_{OE,CU} \neq \mu_{OE,EX}$
- $H2$: The means (μ) of the *random error* are different across both user groups:
 $H2_0 : \mu_{RE,CU} = \mu_{RE,EX}, H2_A : \mu_{RE,CU} \neq \mu_{RE,EX}$
- $H3$: The means (μ) of the *systematic error* are different across both user groups:
 $H3_0 : \mu_{SE,CU} = \mu_{SE,EX}, H3_A : \mu_{SE,CU} \neq \mu_{SE,EX}$

6.5.2.3 Ground Truth Data Set

To obtain ground truth labels for determining error rates of individual labelers, high-quality labels were necessary. We created a data set based on SET_CITY_1 and SET_CITY_2, which were collected using the Twitter Search API (see Section 3.2.1). On these sets, we applied the incident keyword filtering presented in Section 3.2.2. From the resulting set, we randomly selected 2,000 tweets containing at least one incident-related keyword. Those 2,000 tweets were manually labeled by four researchers using an online survey. To assign the final coding, the majority of coders had to agree on a label. Labels for which no agreement was achieved were discussed and relabeled in a group discussion. The final labeled data set used for this evaluation is as follows:

- **SET_1_GT**: 2,000 tweets (309 car incidents, 328 fire incidents, 334 shooting incidents, and 1,029 related to no incident or other type of incident)

6.5.2.4 Evaluation Design, Method, and Procedure

For conducting our evaluation, we used a within-subject design with repeated measures. We chose the user type as the independent variable and error rate, defined as the amount of incorrect annotations, as the dependent variable. Each participant of each type was asked to select a *single* annotation for *all* 2,000 tweets of Set SET_1_GT. As annotations, the incident types presented in Section 3.2.1 were chosen: "Car incident", "Fire incident", "Shooting incident", and "Not incident related or other type of incident".

To test the hypotheses, we created an online survey to conduct the annotation according to the incident types. Participants were invited to participate in the ex-

periment via word of mouth and could execute the annotation tasks using their own computers; it was possible to interrupt and resume the tasks at any time. A start page introduced participants about their task. Then participants were asked to assign a label to tweet. The experiment was fully counterbalanced using a Latin square design. After completion of the task, participants were asked to fill out a short demographic questionnaire.

Our study was conducted with two user groups: domain experts and nonexperts. Fourteen researchers of SAP AG and TU Darmstadt participated in the study. Eight participants were nonexperts with no or low experience in the crisis management domain, and six users were domain experts with more than two years of research experience in the emergency management domain.

Each tweet was labeled by at least three nonexperts and at least two domain experts. Based on the results, we calculated the systematic, random, and overall errors using the ground truth labels in SET_1_GT (see 7.5). Of these, 1,793 tweets received a unique annotation in the ground truth data, and another 207 fell into the multi-annotation category and were ambiguous. The random error is calculated for all tweets with unique annotations and the systematic for all ambiguous tweets. The overall error is calculated for all 2,000 tweets.

6.5.2.5 Results

We first found that the overall error of each user group is 7.2% for nonexperts and 7.9% for domain experts, respectively. This shows that the labeling quality of tweets is rather high. However, the systematic error is high with more than 41%. This is a result of selecting a single label for tweets for which multiple labels would be appropriate. Without further situational information, this selection seems to be difficult, even for domain experts.

For evaluating our hypothesis, we first confirmed normal distribution for all error types and both user groups using the Anderson-Darling as well as the Shapiro-Wilk normality test. Furthermore, we conducted a two-sample F-test for variances to verify same variances for all combinations with $p < 0.01$. For each combination we conducted the two-sample t-test assuming equal variances. For all combinations, the null hypotheses could not be rejected with $p < 0.01$. Thus, for all error types, we cannot assume a difference between both user groups. This means that in our study, there is no conceivable difference between domain experts and nonexperts.

One simple reason for this missing difference might be found in the nature of social media data. Tweets are rather short; the amount of available information per tweet is limited. Thus, the complexity of the information is low, and it is possible to understand the content even as a nonexpert. Furthermore, as tweets are sent by lots

Table 16.: Results for the overall error, random error, and systematic error (μ =mean, σ =standard deviation).

	Overall Error		Random Error		Systematic Error	
	Nonexpert	Expert	Nonexpert	Expert	Nonexpert	Expert
μ	0.0729	0.0790	0.0338	0.0323	0.4106	0.4435
σ	0.0006	0.0003	0.0006	0.0002	0.0053	0.0077

of different individuals, the amount of domain-specific terms could be rather low compared with specialized texts.

6.5.2.6 Study 2: Summary

With this evaluation, we showed that human-based classification seems to be an appropriate means for annotating data sets of user-generated content in the domain of emergency management. However, to further reduce error rates, multiple annotations should be provided at once for each information item. Furthermore, we showed that there seems to be no significant difference in labeling quality between domain experts and nonexperts for this annotation task. Thus, nonexperts are suitable for human-based classification.

6.6 Conclusion

In this chapter, we presented the *human-based classification and aggregation* step as part of the framework. We showed how crowdsourcing can be used to infer the thematic dimension of an incident [R1] as well as to aggregate information related to the same incident. Furthermore, we presented how crowdsourcing can be used to collect incident-related information. As a result of this processing step, aggregated and classified information can directly be provided to a decision maker or can be used by the machine-based classification step.

In this chapter, we made the following contributions:

- We presented a general approach for applying crowdsourcing to *manually* classify and aggregate user-generated content according to the information need of the command staff in emergency management. We introduced a combination of participatory human-centered sensing for actively gathering information as well as human-based classification and aggregation for filtering user-generated content. With our approach, we are able to manually determine the thematic dimension of an information item. In contrast to previous approaches in the emergency management domain, our approach relies on both crowdsourcing

and different mechanisms for the automatic processing of user-generated content.

- In an evaluation with 23 participants, we showed that participatory human-centered sensing and human-based classification and aggregation are indeed valuable for the command staff. We found that our approach would contribute valuable situational information for daily decision making. Furthermore, we underlined the importance of retrieving additional metadata such as spatial and temporal information from user-generated content.
- In a second study, we evaluated the quality when human-based classification is applied on social media data as manual annotations are prone to errors. We showed that this approach can be used for annotating data with sufficiently high quality. Furthermore, in the evaluation, we could not verify a difference in annotation quality between nonexperts and domain experts. This is an important finding when classifications are needed in the emergency management domain as typical crowdsourcing environments rely on nonexperts as annotators.

For future work, one could follow the approach of related works that augment the initial information base with additional background knowledge, for instance, by providing definitions for certain domain-specific terms. Furthermore, the information base can be extended by providing additional context information for the objects mentioned in the information stream (e.g., indicating other objects next to the incident that require protection).

This augmentation might work as follows: The decision maker may ask a question such as "Is there a fire at university?". This question explicitly refers to the University of Darmstadt, because the spatial extent was defined by the decision maker, although "Darmstadt" is not explicitly mentioned. A corresponding relevant information item to be identified as relevant is a message such as "It is burning at TUD?", where "TUD" is the commonly used abbreviation of "Technische Universität Darmstadt", which is the official name of the Darmstadt university. A tagging engine such as Spotlight (see Chapter 4) may now annotate the named entity mention "university" with the DBpedia category "dbpedia-owl:University", as well as the string "TUD" with the DBpedia entity `dbpedia:Technische_Universität_Darmstadt`. Since there is a link (i.e., `rdf:type`) between the latter two, a similarity score can be computed between the question and the information item in question, e.g., by counting the number of traversed links in LOD. Sorting the information objects by those similarity scores leads to a preclassified collection of information, which has a higher relevance according to a question than other information items. This augmentation process is shown in Figure 28.

As crowdsourcing has limited applicability when it comes to annotating large amount of data, in the next sections, we present our approaches for applying au-

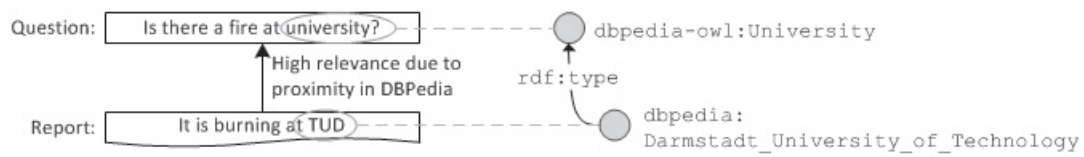


Figure 28.: Example of automatically mapping incident reports to questions based on named entities.

automatic processing methods to infer the thematic dimension of an incident. Furthermore, we show how to automatically aggregate information related to the same event.

7 Machine-Based Classification of User-Generated Content

In the last chapter, we showed how to manually classify the type of incident mentioned in user-generated content (see Figure 29). However, the applicability of crowdsourcing is limited when real-time filtering is needed. One reason for this is that recruiting a critical mass is not always feasible. Furthermore, classifying social media is time-consuming. Thus, automatic approaches for classifying social media data are a necessity. For the automatic classification of large amounts of data, we present a highly precise *and* generalizable approach for inferring the thematic dimension of a tweet (R[1]). As a result of our approach, we are able to classify a large amount of user-generated content with respect to the incident type.

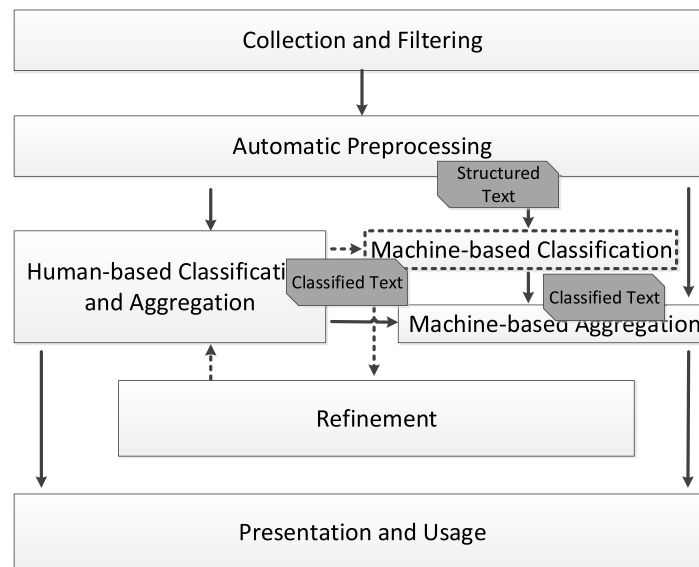


Figure 29.: Machine-based classification as a step in the framework.

As a first research question, we tackle the problem of how to build a supervised learning approach with high classification performance to automatically infer the thematic dimension. Accurate classifications are crucial as misclassification might lead to missing important situational information. This is much more difficult compared with information related to large-scale incidents, especially for classifying information related to small-scale incidents. For the latter one, a system can be optimized for precision, whereas recall is not problematic due to the amount of information that is available for the incident. In contrast, detecting small-scale incidents imposes much stricter demands on *both* precision and recall.

Currently, there are only few systematic evaluations of features for classifying incident-related tweets. Thus, we first conduct feature engineering and evaluate which features are valuable for this classification task. This is subsumed under the following question:

- *How can user-generated content regarding (small-scale) incidents be automatically and precisely classified?*

In an evaluation, we present the result of a comprehensive feature selection approach showing the most valuable features for this task. As a result of this, a classifier is built that precisely classifies the thematic dimension of user-generated content according to the three incident types we defined in Section 3.2, which are "Car incident", "Fire incident", "Shooting incident", and "No incident or other type of incident".

As a second question, we deal with the problem of robustness (i.e., the generalizability of a model to different data sets). This is important as current supervised learning approaches for incident classification focus on regionally restricted data sets such as data from only one city. The important fact that social media data varies considerably across different cities is often neglected; thus, the expectation that a classifier trained on data from one city works precisely on a different city may not hold true.

This is because such a type of text has special properties compared with structured textual information. First, in many social networks, the length of each posting is limited; thus, short messages consist of only few phrases or sentences. Second, social media is inherently unstructured as users tend to use abbreviations or nonstandard vocabulary. This is even increased because of the diversity of authorship; thus, many different styles of writing can be found. Third, named entities used in texts are likely to be related to the location where a text was created or contain certain topics; thus, when the classifier relies on named entities that are unique in the given city such as street names, place names, etc., it will not work well on other cities where these are not used.

To deal with this problem, the data of each city usually needs to be labeled, which is costly and time-consuming. Thus, we also deal with the problem of how to create a model that performs well not only for one city but also for a second city. The question related to this is as follows:

- *How can a general classification model that works well on different data sets be built?*

We tackle this problem by creating a generalized model using training information in the form of social media data collected in one city to classify data that stem from a different city. We introduce a novel approach called *semantic abstraction*, which helps to enrich the feature space with generalized not city-specific features so that they become city-independent. To do so, we use the named entity and temporal ex-

pression recognition presented in Chapter 4 to introduce abstract features based on the occurrence of location and temporal mentions. On the other hand, background information provided by LOD is used to obtain new features that are universally applicable. This is done by extracting named entities and enhancing the feature space with the direct types and categories of the entities at hand. An evaluation of the generalizability shows that the novel approach for semantic abstraction can improve classification results when trained and tested on one data set. Likewise, classification performance is significantly better when a classifier is trained on one city and applied on a different one.

In summary, the contributions of this chapter are the following:

- We propose a set of features that are most suitable for classifying the type of incident in user-generated content.
- We validate the performance of the best feature combination on different tweet data sets and show that we are able to classify the incident type with an F-measure of more than 90%.
- We introduce the novel concept of semantic abstraction, which allows the creation of features that are not city-specific and support training a generalized model. In comparison with related work, our approach combines several approaches for semantic abstraction.
- We evaluate semantic abstraction on tweets that stem from five different cities, showing that it is indeed valuable to overcome city-specific features (i.e., for training a generalized model).

In Section 7.1, we first introduce supervised learning and classifiers used for our evaluation. In Section 7.2 related approaches are presented. Our approach of semantic abstract as well as our automatic classification pipeline is shown in Section 7.3 followed by a description of a prototypical implementation for an emergency management system in Section 7.4. Next, we present the results of our evaluation (see Section 7.5). Finally, we close with a conclusion and future work (cf. Section 7.6).

Parts of this chapter appeared in [201, 197, 154].

7.1 Background

As we want to classify textual user-generated content regarding incidents, we focus on machine learning-based text classification. In general, a text classification problem is formally defined as follows [145]:

Given a document (in the remaining referred to as "instance") $d \in X$ with X as the document space, a set of classes (also called "labels") $C = c_1, c_2, \dots$, and a training set

D of labeled documents (d, c) , we want to learn a classifier γ which maps documents to classes $\gamma : X \rightarrow C$. As labeled training data is provided for training, this type of learning is known as "supervised learning". The trained classifier results in a model that can be applied on unlabeled documents.

According to this mapping, a document is assigned to exactly one class. For instance, if we want to divide tweets into the classes "incident related" and "not incident related", we have a two-class classification task. However, there are dependencies and interconnections in the data that can be detected and exploited in order to obtain additional useful information or just better classification performance. This approach is known as multi-label classification. In multi-label classification j different classifiers γ_i are trained, each returning either c_j or a class vector \bar{c}_j .

For training a classifier, documents need to be converted into a manageable representation [73] (i.e., a feature vector). A feature (or attribute) can be regarded as one dimension in the feature space and is taken into account when making decisions in classification problems.

In the following, we present three different known classifiers that are used throughout this dissertation. We present a (multinomial) Naive Bayes (NB) classifier and Support Vector Machines (SVM) as both are easily updatable with new instances and showed good performance for text classification [113]. As a third classifier, we introduce the Repeated Incremental Pruning to Produce Error Reduction (RIPPER) rule learner. This classifier is mainly used to create human readable rules so that we are able to get a better understanding of our feature sets.

Naive Bayes Classifier

Naive Bayes (NB) is a probabilistic classifier that calculates the probability $p(d|c)$ that a document d belongs to a class c [145]. Given the document representation as a feature vector $d = (f_1, f_2, \dots)$ the probability $p(d|c)$ is computed as shown in Equation 10, with $P(f_i|c)$ as the conditional probability of a feature f_i occurring in a document of a class.

$$p(d|c) = P(c) \prod_i P(f_i|c) \quad (10)$$

$P(f_i|c)$ is the probability of how much a feature f_i contributes for predicting the correct class c . This is multiplied with the prior probability $P(c)$ of a document occurring in class c . The class with the prior probability is chosen if the features do not provide enough evidence for deciding for one class over another. Finally, with the NB approach, we find the most likely class a document belongs to.

One important aspect of NB is that it assumes conditional independence. However, for text data, this does not hold true. Nevertheless, it has been shown that NB performs well even with dependent data [145].

Support Vector Machines

Another classifier commonly used for text classification is an SVM [55]. As shown in Figure 30, SVMs calculate a decision boundary between classes so that they are maximally far from any point in the training set.

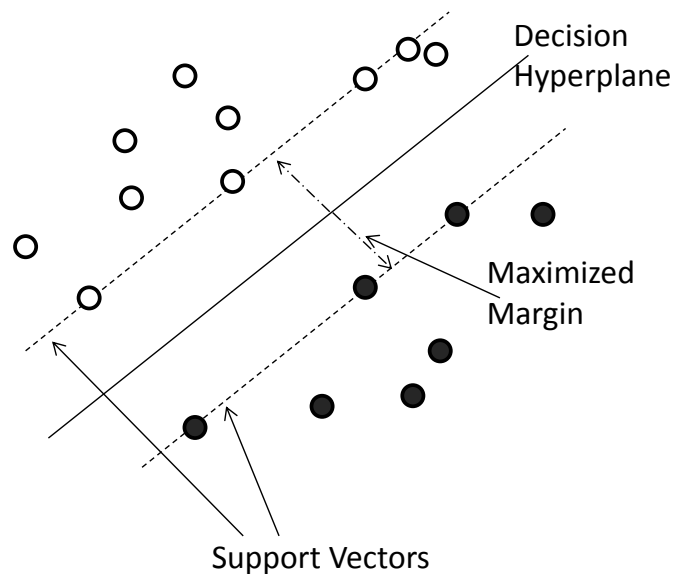


Figure 30.: Example of decision hyperplane, margin, and support vectors for a binary linear classifier.

This decision boundary is called hyperplane and separates both sets of instances. The distance from the decision hyperplane to the closest data point is called margin. For finding the optimal results, this margin is maximized. As shown in the figure, the decision function is specified by a subset of the data. These subsets are called the support vectors.

The example shown in Figure 30 is representative for two-class classifiers. However, we also apply multi-class classification. For applying an SVM as a multi-class classifier, it is common practice to build a "one-versus-all" classifier. This classifier selects the class that has the greatest margin.

In practice, there are cases when classes are not linearly separable. In these cases, a higher order function may be applied to split the data set. This can be done using a non-linear SVM (or the so-called kernel trick; see [145] for more details on this aspect). Nevertheless, for our evaluations, we use linear kernel, which was shown as comparable to non-linear kernels for an incident classification task (see [38],

[186]). Furthermore, it has been shown that for a large number of features and a low number of instances, a linear kernel is comparable to a non-linear one [101].

As another important aspect of SVMs, their effectiveness depends on the selection of various parameters. In our case, the standard linear SVM model can be extended to ignore noise by paying a cost for a misclassified example [145]. This is implemented by a slack variable c . The slack variable measures the degree of misclassification allowed for each instance. Thus, increasing c leads to less misclassified examples on the training data. However, the model becomes a less general model; this means it is *overfitted* to the training data. In our evaluation, we evaluate the best parameter settings for c whenever an SVM classifier is used as our goal is to find the best classifier.

Decision Rule Classifiers

Models trained with the classifiers presented so far cannot be easily understood by humans. Thus, we also make use of a symbolic classifier that allows the creation of human readable rules. In general, a classifier is built by regarding each training document as a clause [73]:

$$f_1 \wedge f_2 \wedge \dots \wedge f_n \rightarrow c \quad (11)$$

with f_i as the features of the document and c its class. A rule learner then generalizes these rules until the best rules that correctly classify all training examples according to some optimality criterion are found.

We use the RIPPER rule learner [49] as one implementation of decision rule classifiers. RIPPER creates rules in a greedy manner by first adding rules so that all positive instances are covered and then adding conditions so that no negative instance is covered. Overfitting is avoided by optimizing the rule set posterior so it is reduced in size and is likely to fit the training data.

In a multi-class case, the number of instances belonging to a class is counted. Then rules are learned on the smallest class first, treating the rest as negative class. This procedure is repeated with the next smallest class and so on.

7.2 Related Work

As shown in Chapter 6.2, several approaches apply (keyword-based) filtering or machine learning for detecting user-generated content relevant for emergency management. In this related work section, we focus on approaches that try to classify incident-related user-generated content.

We first give an overview of approaches using supervised learning techniques for classifying user-generated content related to large-scale and small-scale incidents. We especially take a look at features that were used for this classification task. Furthermore, we present related approaches that try to learn a generalized model for classifying user-generated content.

7.2.1 Incident Type Classification

In the following, we give an overview of approaches for the classification of incident-related user-generated content. Related approaches are differentiated with respect to the corpus used for incident type classification and the scale of the incident type addressed. Furthermore, approaches differ in the classifier that is used and the number of classes that are detected. Also, different feature groups are used. Finally, we analyze related works according to their approaches of training a generalized classifier (i.e., according to their use of semantic abstraction). An overview of related approaches is given in Table 17 and Table 18.

Table 17.: Overview of related approaches for incident type classification.

Approach	Corpus	Scale of Incident		Classifier	#Classes
		Large	Small		
Sakaki and Okazaki [192]	Tweets	x		SVM	2
Becker et al. [23]	Cluster	x		SVM	2
Hua et al. [103]	Cluster		x	SVM	2
Agarwal et al. [5]	Tweets		x	NB, SVM	2
Wanichayapong et al. [240]	Tweets		x	Keyw.	2
Li et al. [137]	Tweets		x	unknown	2
Carvalho and Rossetti [38]	Tweets		x	SVM	2
Robert Power [186]	Tweets		x	Keyw., SVM	2
Walther and Kaiser [238]	Cluster		x	JRip	2
Karimi et al. [119]	Tweets		x	SVM	6
Our Approach	Tweets		x	Keyw., NB, SVM	2/4

Sakaki et al. [192] used an SVM classifier to detect earthquakes as a type of large-scale incident. The SVM was trained using three features extracted from tweets specifically referring to earthquakes: the number of words occurring in the tweets, statistical features (the number of words in a tweet and the position of keywords), and word context features (the words before and after the earthquake-related keyword). They used a data set of 597 earthquake-related tweets and showed that their approach has a precision of 63.64% and recall of 87.50% for differentiating earthquake-related tweets and not related tweets.

Table 18.: Overview of related approaches for incident type classification with respect to the used feature groups as well as according to the use of semantic abstraction. (Named Entities = NEs)

Approach	N-Grams	#NEs	#URLs	TF-IDF	Twitter	Other	Sem. Abstr.
Sakaki and Okazaki [192]	x					Contextual	
Becker et al. [23]	(x)				x	Buzzy Terms	
Hua et al. [103]				x			
Agarwal et al. [5]	x	x	x				
Wanichayapong et al. [240]							
Li et al. [137]		x			x		
Carvalho and Rossetti [38]	x						
Robert Power [186]	x				x		
Walther and Kaisser [238]						Sentiment	(x)
Karimi et al. [119]	x				x		
Our Approach	x	x	(x)	x		Sentiment	x

Becker, Naaman, and Gravano [23] presented a system for event detection. Based on cosine similarity of the TF-IDF scores of each tweet to a cluster, a preclustering of tweets is performed [22]. Afterward, each cluster is assigned a label whether it is an event cluster. For this, they use a combination of temporal (i.e., prominent terms), social (i.e., interaction such as retweets and replies), topical features (i.e., common terms), and Twitter-centric features (i.e., presence of hashtags). Based on this, a SVM classifier is trained. An evaluation was conducted on 374 manually annotated event clusters consisting of tweets from New York City and showed an F1 score of 83.7%.

The previous approaches focus on large-scale incidents. In contrast, other state-of-the-art approaches focus on the detection of small-scale incidents.

Hua et al. [103] presented STED, a system for small-scale event detection. Like Becker et al. [23], they apply text classification for classifying preclustered tweets. Compared with other approaches, named entities are discarded before calculating TF-IDF scores, which are the only features used in their approach. An SVM was trained for a specific event type and applied on the clusters. The approach was tested on (an undefined number of) tweets collected in Latin America and shows a precision of 72% and recall of 74% for classifying the clusters.

Agarwal et al. [5] proposed an approach for classifying tweets related to a fire in a factory. As a first step, their system detects incident-related messages using a combination of an NB and an SVM classifier. As features, they use the number of occurrences of certain named entities such as locations, organizations, or persons

that are extracted using the Stanford NER toolkit. Furthermore, the occurrence of numbers and URLs is used as a feature. Also, word occurrences remaining after stopwords filtering are used. The approach was tested on 1,400 tweets and shows that they are able to detect tweets related to factory fires with up to 80% accuracy. Furthermore, they showed that NB outperforms the SVM classifier, which could be because of an untuned SVM.

Wanichayapong et al. [240] focused on extracting traffic information in tweets from Thailand. Their approach mainly relied on a dictionary-based approach. First, tweets are prefiltered using traffic-related keywords. Second, traffic-related keywords in combination with location-related keywords are used to classify traffic tweets. An evaluation of 1,249 Twitter messages shows that this simple approach is able to give a precision of 91.39% and a recall of 87.53%.

Li et al. [137] introduced a system for the searching and visualization of tweets related to small-scale incidents based on keyword, spatial, and temporal filtering. Compared with other approaches, they iteratively refine a keyword-based search for retrieving a higher number of incident-related tweets. Based on these tweets a (not named) classifier is built upon text features and Twitter-specific features, such as hashtags, @-mentions, URLs, and the number of spatial and temporal mentions. They report an accuracy of 80% for detecting incident-related tweets, although they do not provide any information about their evaluation approach and the classifier used.

Carvalo, Sarmiento, and Rossetti [38] evaluated an automatic classification of traffic-related tweets. Compared with other works, they conducted no initial labeling but used a set of tweets from official sources as ground truth data. An SVM classifier was trained based on this and (manually) evaluated on the rest of the tweets. As features, they used simple word unigrams, after stopwords and punctuation removal. Furthermore, they showed that an SVM with linear kernels gives the same performance as other kernels. Finally, they achieved an F-measure of approximately 23%.

Power, Robinson, and Ratcliffe [186] analyzed how to detect tweets related to fire incidents. In a preliminary evaluation, they showed that a simple keyword-based approach using the observed frequency of a word compared with historical frequency gave an accuracy of 48%. In a second evaluation, an SVM with a linear kernel function was trained. They analyzed several feature combinations based on the number of words, user mention count, hashtag count, hyperlink count, unigram occurrences, and bigram occurrences. They found that a combination of both unigram occurrences and user mention count gave the highest performance with an F1 score of 83.1% on 794 tweets.

Walther and Kaisser [238] presented an approach for small-scale event detection. However, their goal was not to annotate a single tweet but to identify an event based on a set of tweets. As textual features, they used sentiment features, bi-

nary weighting of most frequent terms, and several dictionary-based feature groups. From these, they used a semantic dictionary, which contains a list of terms related to higher-level event categories such as "sport events". Their approach has been evaluated with 1,000 manually labeled events (they do not provide the overall number of tweets) and evaluated using JRip. They showed that they achieve a precision of 85.8% and a recall of 85.6% for classifying the cluster of tweets.

Karimi, Yin, and Paris [119] tried to classify tweets according to six incident type classes. They relied on unigrams and bigrams as well as Twitter-specific features such as hashtags and @-mentions. The approach was evaluated on 5,747 tweets and showed an accuracy of up to 90% when using 90% of the data as a training set. Precision and recall were not provided. However, compared with other approaches, they did not conduct cross validation but time-split evaluation. Thus, older data is used for training to deal with the dynamism of user-generated content. Furthermore, they showed that by using an SVM classifier, the best results could be achieved.

7.2.2 Approaches Training a Generalized Model For Classifying User-Generated Content

Using external knowledge sources such as DBpedia as well as information about named entities was proposed in related work several times [171, 98]. In the following, we present approaches that are related to our semantic abstraction approach (see Table 19).

Table 19.: Overview of approaches that are related to our semantic abstraction.

Approach	Classification	Semantic Abstraction			Knowledge Base
		Location	Temporal	NEs	
Saif et al. [191]	Sentiment	(x)		x	DBpedia
Cano et al. [37]	Topic	(x)		x	DBpedia, Freebase
Song et al. [217]	Topic	(x)		x	WordNet, Freebase, Wikipedia, Probase
Xu and Oard [248]	Topic	(x)		x	Wikipedia
Muñoz García et al. [157]	Topic	(x)		x	DBpedia
Our Approach	Incident Type	x	x	x	DBpedia, internal

Saif et al. [191] showed that adding the semantic concept for a named entity is valuable for sentiment analysis on tweets. They used the concept tagging part of the AlchemyAPI⁴³ to extract one concept for each named entity in a tweet and use it as a feature. For instance, the concept "President" is derived for "Barack Obama". Their results show that semantic abstraction works well for very large data sets with

⁴³ <http://www.alchemyapi.com/api/concept-tagging/> [Accessed: 22.05.2013]

a multitude of topics, but not on small data sets. Compared with their work, our approach makes use of multiple types and categories extracted for a named entity, providing us with a much richer set of background information.

Cano et al. [37] proposed a framework for topic classification, which uses Linked Data derived from DBpedia and Freebase for extracting semantic features. They compared the approach with a baseline comprising TF-IDF scores for word-unigrams, concepts extracted using the OpenCalais API⁴⁴, and Part-of-Speech features. For comparison, an SVM classifier is used. In an evaluation of 10K tweets, they show that semantic features are indeed useful compared with the baseline approach. They outline that their current approach does not resolve abbreviations; thus, it is prone to miss named entities.

Song et al. [217] also proposed an approach that makes use of concepts derived for instances of tweets using external knowledge databases for topic clustering. In an evaluation, they compared different knowledge sources such as WordNet⁴⁵, Freebase⁴⁶, Wikipedia⁴⁷, and Probase⁴⁸. They performed a k-means clustering on tweets. They showed that using conceptualized features, it is possible to outperform a plain bag-of-words approach. Xu and Oard [248] followed a similar approach for topic clustering, but focused on Wikipedia as a knowledge source. Information from Wikipedia is used as additional features to identify topics for tweets. They also showed an improvement compared with not using this information.

Muñoz García et al. [157] tried to use DBpedia resources for topic detection. Their approach uses Part-of-Speech tagging for detecting nouns. These nouns are inter-linked to DBpedia resources using the Sem4Tags tagger [79]. The approach was evaluated on textual user-generated content derived from sources such as Twitter, Facebook, and Flickr. In an evaluation, they showed that extended features are valuable for topic detection.

The approaches presented so far are examples of how to make use of external knowledge sources for deriving abstract features. Some approaches tried to deal with geographical variations for topic detection and geolocalization of user-generated content using language models [100, 66]. Although these approaches deal with a similar problem, the application of these approaches for a supervised learning problem is limited.

Despite the works focusing on user-generated content, creating a generalized model was investigated in other domains, for instance, in the area of transfer learning [61, 180]. Domain adaptation [58] is also related to our approach. However, where

⁴⁴ <http://www.opencalais.com/calaisAPI> [Accessed: 21.05.2013]

⁴⁵ <http://wordnet.princeton.edu/> [Accessed: 21.05.2013]

⁴⁶ <https://www.freebase.com/> [Accessed: 21.05.2013]

⁴⁷ <https://www.wikipedia.org/> [Accessed: 21.05.2013]

⁴⁸ <http://research.microsoft.com/en-us/projects/probase/> [Accessed: 21.05.2013]

the domains are to a large extent different in domain adaptation, in our setting the domain (i.e., incident type classification of tweets) remains the same, but the input data is subject to change. This means that certain features (i.e., words) are changing from city to city. Therefore, feature augmentation [57] is related to our approach. However, where domain-specific features are simply discarded in regular feature augmentation, our method abstracts them in advance, and then they are used in union with domain-independent features. Another way of adapting domains is structural correspondence learning [30], where shared features are identified, augmented, and used to build classifiers that are applicable in both domains. The main difference is that the shared features that are then used have to be present. However, we instead create these shared features based on existing ones by the proposed semantic abstraction.

7.2.3 Summary

In this related work section, we showed that many related approaches focus on classifying user-generated content according to the type of incident, either for large-scale or small-scale incidents. The approaches are directly applied on tweets or on already-clustered sets of tweets. We found that related approaches mostly rely on NB and SVMs as classifiers for this task. Also, and in contrast to our approach, mostly two classes are differentiated. For this classification task, many different feature sets were proposed, comprising general word-based features such as the number of word occurrences, TF-IDF scores, and sentiment features. Furthermore, platform-specific features were also used.

Although [238] showed how to use conceptual information about named entities, training a generalized model for incident type classification was not applied so far. Also, though several knowledge bases were evaluated in related works, mostly one source is used for the respective classification tasks. In contrast to this, we make use of DBpedia in combination with two internal approaches for extracting and generalizing location mentions and temporal expressions. Furthermore, and most importantly, none of the existing approaches were evaluated on data that stem from multiple cities.

7.3 Approach

In the following section, we present our approach for classifying the type of incident mentioned in user-generated content. We first present the general pipeline for finding an optimal set of features for this classification task. Second, we introduce the novel approach of semantic abstraction, which allows us to train a generalizable model.

In Figure 31, the pipeline for automatic incident type classification is shown. The pipeline is divided into three parts and follows the standard approach for feature generation (see [245]). In the pipeline, we make use of feature groups that are commonly used in related works. Furthermore, we include our approach of semantic abstraction prior to generating other feature groups.

1. In the first step, social media data is collected, for instance, as part of our collection and filtering step (see Chapter 3) or the human-based classification and aggregation step (see Chapter 6). As a result of this, several instances of documents are retrieved. These instances of unstructured data are preprocessed to allow feature generation (see Chapter 4).
2. As a next step, semantic abstraction takes place. Temporal expressions and location mentions are identified and replaced with common tokens. Likewise, named entities are mapped to the corresponding entities in LOD (see Chapter 4).
3. Subsequently, several features that can be used for training a classifier for incident type classification are generated. The selected feature groups are based on the features commonly used in related work. However, we do not use Twitter-specific features as this would restrict our approach to Twitter data.

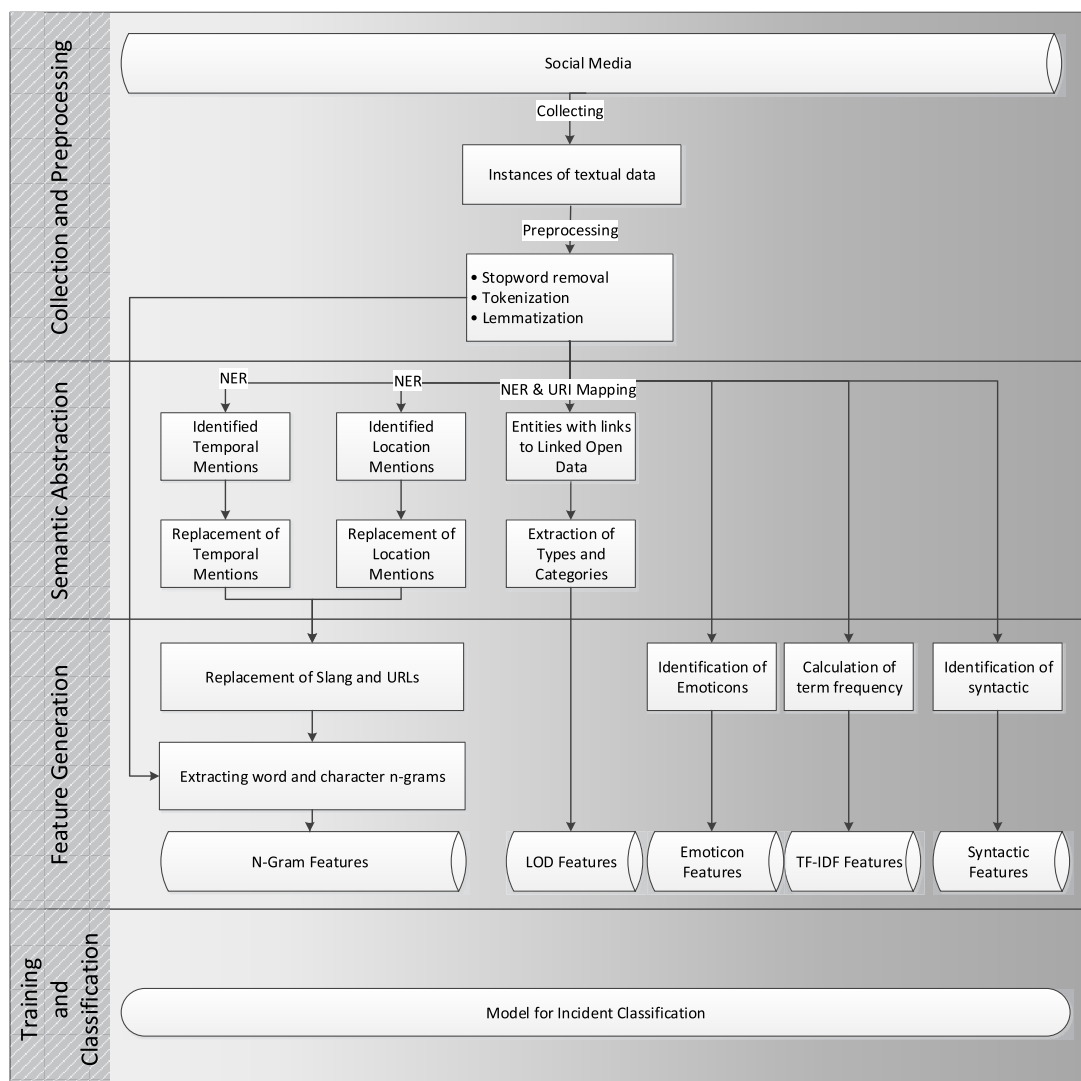
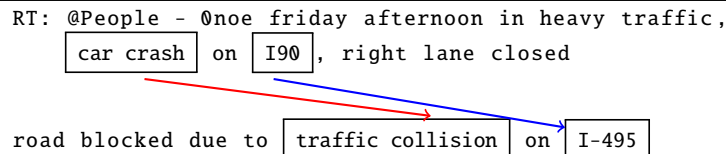


Figure 31.: Pipeline for automatic incident type classification showing semantic abstraction, feature generation, and training steps.

7.3.1 Semantic Abstraction

To illustrate the idea behind semantic abstraction, two tweets in Listings 7.1 are shown, which both describe an incident:

Listing 7.1: Semantic similarity between two example tweets.



```
RT: @People - Onoe friday afternoon in heavy traffic,  
car crash on I90, right lane closed  
  
road blocked due to traffic collision on I-495
```

Though both tweets describe an incident, this similarity between both texts is not easily identified. Nevertheless, both tweets consist of entities that might describe the same thing with different wording. In this example, "accident" and "car collision" are similar expressions for the same type of event. Furthermore, "I90" and "I-495" are both names of streets. With simple text similarity approaches, it is not easily possible to make use of this semantic similarity, although, as we show, it is valuable for classifying both tweets.

Because of the special properties of social media data, generalization by only using standard bag of word features is assumed to be difficult. Thus, we developed a novel approach called semantic abstraction. Semantic abstraction allows enriching the feature space with features that can be used to overcome city- and text-specific properties.

Before generating features using semantic abstraction, the individual NER steps presented in Chapter 4 are performed on the unprocessed text: (1) Named Entity Recognition and Replacement using Linked Open Data, (2) Location Mention Extraction and Replacement, and (3) Temporal Expression Recognition and Normalization on Unstructured Text. As a result of these steps, we extract URIs for named entities with links to the corresponding entities in LOD as well as location mentions and temporal expressions in the text. Based on this, semantic abstraction can be performed.

In the following, we present three approaches for generating feature groups using semantic abstraction.

(LOD) Linked Open Data Feature Group: As a first approach, we use LOD as a source of interlinked information about entities. As presented in Chapter 4, we make use of two relations that are present in DBpedia. First, we use type relationships, which are the URIs of a class or the resources describing the class of a named entity. Second, we use categories (i.e., the subject relationship) that relate to one or many resources describing the topic of a named entity.

To enrich our feature space with LOD features, we use the RapidMiner Linked Open Data extension [171] to map named entities to corresponding URIs in DBpedia. To do this, a process was modeled in RapidMiner as shown in Figure 32.

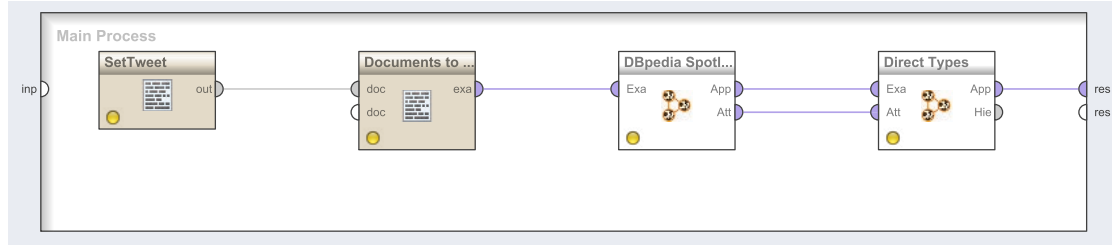


Figure 32.: RapidMiner process for extracting direct types of entities.

The process consists of four steps. First, a data set of tweets is generated. Next, the extension proceeds by recognizing entities based on DBpedia Spotlight [155] to get URIs of the detected named entities. Third, the URIs of detected entities are used to finally extract the direct types. These types are contained in the *TYPE* feature group, which is part of the LOD feature group.

To identify all subject relationships, we use the identified URIs and extract all categories. For this, the following SPARQL query is used:

Listing 7.2: SPARQL query for extracting categories for a named entity using its URI.

```
SELECT DISTINCT ?y WHERE {<" + URI + "> <http://purl.org
/dc/terms/subject> ?y .}
```

The query is applied for all URIs and extracts the corresponding subject (i.e., the categories) for each entity. This results in the *CATEGORIES* feature group, which is also part of the LOD feature group. For instance, for the following tweet, the types and categories shown in Table 20 can be extracted.

Car crash on Interstate 90, everything is on fire.


Table 20.: Extracted Types and Categories for Tweet "Car crash on Interstate 90, everything is on fire".

Types	Categories
dbpedia-owl:ArchitecturalStructure	category:Interstate_Highway_System
dbpedia-owl:Infrastructure	category:Interstate_Highways_in_Indiana
dbpedia-owl:Place	category:Interstate_Highways_in_Wyoming
dbpedia-owl:Road	category:Interstate_90
dbpedia-owl:RouteOfTransportation	category:Interstate_Highways_in_Massachusetts

Based on the extracted types and categories for each tweet, the number of occurrences is modeled as a numeric feature. Thus, each of the entries in Table 20 receives a number stating how often the type or category could be extracted for the example tweet. As a result of this approach, we are able to abstract named entities in unstructured texts to abstract features, which, as we show, allow training a much more generalized model.

(LOC) Location Mention Feature Group: As a second approach, we replace location mentions with a common token. For this, we use our location mention extraction approach (see Section 4.2.2) to detect the corresponding named entities in the text and replace location mentions in the unprocessed tweet texts as shown in Listing 7.3.

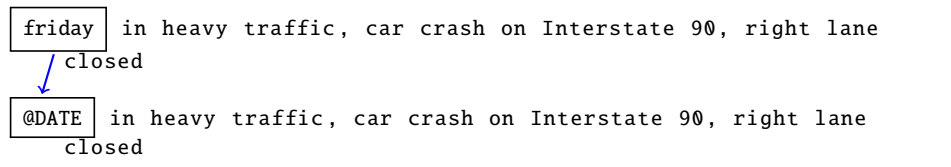
Listing 7.3: Example of replacement of location mention with common token.

friday in heavy traffic, car crash on	Interstate 90	, right lane
closed		
		
friday in heavy traffic, car crash on	ProperLOC	, right lane closed

The location mention extraction is based on a retrained named entity recognizer, which is able to annotate location mentions in tweets as shown in the listing. Based on this, mentions are detected and replaced with a general annotation "ProperLOC". We also detect common location mentions such as "home", "office", or "school" and replace them with a general annotation "CommonLOC".

Based on the text containing the replaced token, further features can be generated. For instance, the replaced location mentions are represented as TF-IDF features or as n-grams. Furthermore, we count the number of location mentions in a tweet. Finally, this results in a group of features for location mentions.

(TEMP) Temporal Expression Feature Group: As a third approach, the same mechanism applied for location mentions, is also applied to the temporal expressions. Thus, our approach for temporal expression extraction (see Section 4.2.4) is applied. The presented approach relies on the adapted HeidelTime framework for the identification of temporal expressions in texts. Before applying the approach, abbreviations and slang are resolved. Next, detected temporal expressions are replaced with two annotations "DATE" and "TIME" as shown in Listing 7.4. Based on the text containing the replaced token, further features can be generated. Also in this case, the replaced temporal expressions are represented as TF-IDF features or as n-grams.

Listing 7.4: Example of replacement of temporal mention with common token.

We showed how three different approaches can be applied to generate features using semantic abstraction. In the evaluation in Section 7.5, we will show that this approach is indeed valuable for creating a generalized model.

7.3.2 Feature Generation

After finishing the initial preprocessing steps as well as the semantic abstraction, we extract several features that are used for training a classifier. In this section, we give an overview of the different feature groups that are derived and used for finding the best feature combination for incident type classification. As mentioned before, we do not use Twitter-specific features as this would restrict our approach to Twitter data.

Before making use of our data sets, we needed to convert the texts into a structured representation so it could be used for feature generation. Thus, we conducted the common preprocessing such as the removal of stopwords, tokenization, and lemmatization, thus giving us normalized discrete words (*tokens*).

N-Grams

The first and common approach is to use word or character n-grams as features. An n-gram is a sequence of n elements of characters or tokens in a document. A regular "bag of words" (also called unigram or word-1-gram) representation is commonly used; thus, every token is used as one feature (e.g., $d = \{\text{'car'}, \text{'crash'}, \text{'Interstate'}, \text{'DD'}\}$). However, as this approach makes use of a single token, the context of words due to word co-occurrences is lost. Hence, all combinations of n-grams that are consecutive in the document could be used (e.g. $d = \{\text{'car'}, \text{'car_crash'}, \text{'crash'}, \text{'crash_Interstate'}, \dots\}$).

As it is unclear whether word or character n-grams perform best for our classification task, we compare character-n-grams as well as word-n-grams. Due to the restricted length of social media data, we restrict our evaluations to a maximum of five consecutive tokens or characters.

We showed before that a feature vector needs to be created to train a classifier. Thus, every n-gram could be modeled as binary or frequency-based feature. An approach based on term frequency (TF) enables weighting important words in a document (e.g., $d = \{1, 3, 0, 1\}$). In the TF-based approach, the number of times a word w occurs in document d is calculated:

$$tf(w, d) = |\{w \in d\}| \quad (12)$$

However, binary weighting could also be used on short texts (e.g., $d = \{1, 1, 0, 1\}$). In the binary approach, the term presence (tp) is set if a word w is present in document d :

$$tp(w, d) = \begin{cases} 1 & \text{if } w \in d \\ 0 & \text{if } w \notin d \end{cases} \quad (13)$$

As it is also not clear which weighting scheme performs best, we compare both approaches in our evaluation.

Replacement Strategies

Text shared in social media is inherently unstructured. Users tend to use abbreviations or nonstandard vocabulary in their posted content. Thus, before generating n-grams, social media texts can further be cleaned. To do so, two additional replacement steps can be conducted, which are part of the extended preprocessing pipeline presented in Section 4.2.

First, as abbreviations and slang are commonly used in social media data [183], we replace them with text provided by the Internet Slang Dictionary & Translator⁴⁹. For instance, the token "idk" is replaced with three tokens "I don't know". This way, we avoid too many different tokens as abbreviations are replaced with a common one.

Second, URLs are widely used in user-generated content for referring to external content. These URLs are mostly different; thus, this results in a variety of different URLs that would be added as n-grams. To avoid this, we experiment with replaced URLs to a common token "URL". The replacement is conducted using standard regular expressions.

Syntactic Features

Along with the features directly extracted from the tweet, several syntactic features are expected to improve the performance of our approach. For instance, as the following tweet shows, a repetition of punctuations could be an indicator for a person that is expressing emotions that are the result of an ongoing incident:

⁴⁹ <http://www.noslang.com/> [Accessed: 01.03.2014]

Shots were just fired at the UW bookstore...stay safe everyone!!!

We think that people might tend to use a lot of punctuations, such as exclamation marks and question marks, or a lot of capitalized letters when they are reporting an incident. Thus, we extract the following features:

- The number of "!"
- The number of "?"
- The number of capitalized characters

EMO: Emoticon Feature Group

Emoticons are widely used to express emotions in textual content. Various text classification approaches make use of these, for instance, for sentiment analysis [4, 81]. For incident type classification, they could also be useful as people link emotions with ongoing incidents:

Not even on the highway yet and I'm dead stopped in traffic. :(

For this dissertation, an emoticon library was created based on the suggestion from Agarwal et al. [4]. We extracted a set of 63 emoticons from Wikipedia⁵⁰ and grouped them into the seven categories shown in Table 21. These emoticons are identified in the social media text. The number of occurrences is counted for each category, resulting in seven additional features.

Table 21.: Features and emoticons used for EMO feature group.

happyFace	>:] :-) :) :o) :] :3 :c) :> =] 8) =) :} :^
laughingFace	>:D :-D :D 8-D 8D x-D xD X-D XD =-D =D =-3 =3 8-)
veryHappy	:))
sadFace	>:[:-(- (: :-c :c :-< :< :-[:[:{ >.> <.< >.< :'-) :')
angry	:-
surprise	>:o >:O :-O :O °o° °O° :O o_O o_0 o.O 8-0
disgust	D:< D: D8 D; D= DX v.v D-':

Furthermore, we calculate a simple sentiment score by treating emoticons from "sad", "angry", and "disgust" as negative emotions and "happyFace", "laughingFace", "veryHappy", and "surprise" as positive emotions. Based on these two aggregations, we calculate a simple score, which is added as an additional feature:

$$EmoScore = \sum (posEmoCount + negEmoCount) \begin{cases} posEmoCount = 1 \\ negEmoCount = -1 \end{cases} \quad (14)$$

⁵⁰ http://en.wikipedia.org/wiki/List_of_emoticons. [Accessed: 15.01.2014]

However, there are indeed more sophisticated approaches for calculating sentiment scores as we also showed in our work [202].

TF-IDF Scores and Sum of TF-IDF Scores

Documents describing the same event tend to have a higher similarity compared with documents describing a different event. To express this, token frequencies are represented as their relative occurrence frequency in the document and over the entire corpus. This is known as the vector-space-based TF-inverse document frequency (TF-IDF) approach [6] and is calculated as shown in Equation 15.

$$\text{tf-idf}(w, d, D) = \text{TF}(w, d) * \text{idf}(w, D) = \text{TF}(w, d) * \log \frac{D}{d(w, d)} \quad (15)$$

With this approach, the frequency of a word w (TF) in a document d is weighted with the inverse document frequency (IDF) of a word w in the total number of document in the corpus D . As a result of this, frequent tokens are weighted low, whereas important and more discriminative words are weighted high. We calculate the TF-IDF scores after preprocessing; thus, stopwords are already filtered out.

We use the TF-IDF scores as a static model; thus, tokens that do not occur in the training set are ignored. Hence, new tokens do not receive a weight. Though this might be detrimental for taking new knowledge into account, the opposite approach would result in an overweighting of novel tokens. However, there is current research going on to deal with an incremental TF-IDF model [33].

As a second approach, we follow the idea of using only one similarity score for each instance. This allows for reducing the number of features as only one feature would be created compared with the whole TF-IDF vector. We compute the similarity score as the sum of single TF-IDF scores for each document d :

$$\sum \text{tf-idf}(w, d, D) \quad (16)$$

Semantic Abstraction

As shown before, for the LOC and TEMP feature extraction approach, the resulting features are represented as TF-IDF features and/or are part of the n-grams. Furthermore, the number of location mentions and temporal expressions in a tweet are counted and are present as features. Additionally, for each instance, the extracted types and categories are present as single features with the respective number of occurrences.

7.3.3 Classification

Finally, all extracted features are transformed into a vector representation so they can be used for classification algorithms. Based on the feature vectors, a model for incident type classification is trained.

7.4 Prototypical Realization

In Section 6.4, we presented the initial prototype of the Incident Classifier. Based on the automatic classification approach developed in this chapter, we extended the initial implementation with data retrieved from social media data. Tweets are classified using the machine learning models presented in this chapter and are displayed in the prototype.

In Figure 33, the aggregation of different incident-related information is shown. Furthermore, images that are referenced in social media data are extracted and also displayed in the prototype. For example, in this case, a picture of the incident as well as the number of involved cars is shown. The prototype shows how additional information about an incident might be presented to a decision maker.



Figure 33.: Integration of classified tweets into the Incident Classifier application.

7.5 Evaluation

In the following section, we present several studies showing the performance of our approach for incident type classification. Furthermore, we demonstrate the effectiveness of applying semantic abstraction. First, we give an overview of the data sets and metrics used for our evaluations. Second, we present the results of using a simple keyword-based approach for incident type classification. Third, the results of our feature engineering approach are shown, giving an overview of our best feature set for incident type classification. Also, initial results regarding semantic abstraction are presented. Finally, in the last study, we present the outcomes of analyzing semantic abstraction with respect to training a generalized model.

7.5.1 Data Sets, Metrics, and Methodology

In the following, we present the data sets and metrics used for the evaluation as well as our evaluation design.

Data Sets

For our evaluations, we focused on a two-class as well as a four-class classification task. For the two-class task, we differentiated the following two classes:

- Incident related
- Not incident related

In the four-class task, we chose four classes that match our three incident types defined in Section 3.2.1. Thus, for our machine learning experiment, we focused on three classes consisting of very common and distinct incident types and one neutral class:

- Car incident
- Fire incident
- Shooting incident
- Not incident related or related to other type of incident

As there are no public data sets available for incident type classification, a direct comparison of our approach based on common data sets is not feasible. Thus, we needed to create our own ground truth data.

As ground truth data, we used data sets SET_CITY_1 to SET_CITY_3, which were collected using the Twitter Search API (see Section 3.2.1). As we needed a high-quality ground truth for our experiments, our initial data sets needed to be further

reduced. Thus, we applied the incident keyword filtering presented in Section 3.2.2. From the resulting set, we randomly selected subsets of tweets from the resulting sets containing at least one incident-related keyword.

The tweets resulting from SET_CITY_1 and SET_CITY_2 were manually labeled by at least three researches of our research departments who have experience in emergency management and data labeling. Every tweet was labeled by each researcher. To assign the final coding, the majority of all coders had to agree on a label. In the case of disagreement, issues were resolved in a group discussion. The tweets resulting from SET_CITY_3 were labeled at the crowdsourcing platform CrowdFlower⁵¹. Also in this case, each tweet was labeled by at least three annotators. To assign the final coding, two-thirds of all coders had to agree on a label. In the case of disagreement, the author of this dissertation resolved open issues.

For our evaluation, we created two data sets containing tweets from the cities of Memphis and Seattle to evaluate the general applicability of our classifier as well as to evaluate our semantic abstraction approach. The class distributions are the following (see also Figure 34):

- **4-CLASSES:** 2,000 tweets (328 fire, 309 crash, 334 shooting, 1,029 no incident or other type of incident)
- **2-CLASSES:** 3,286 tweets (1161 incident related, 2125 no incident or other type of incident)

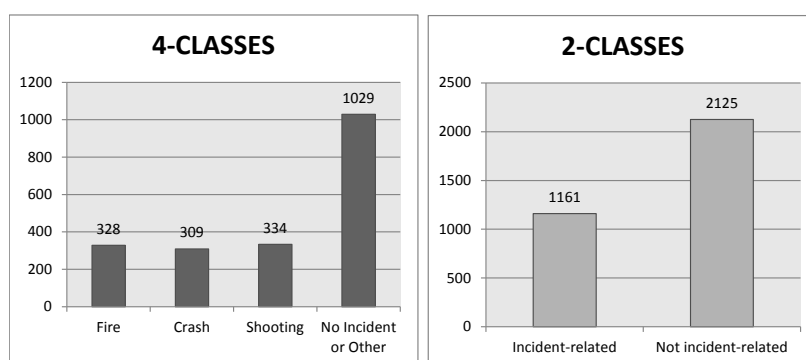
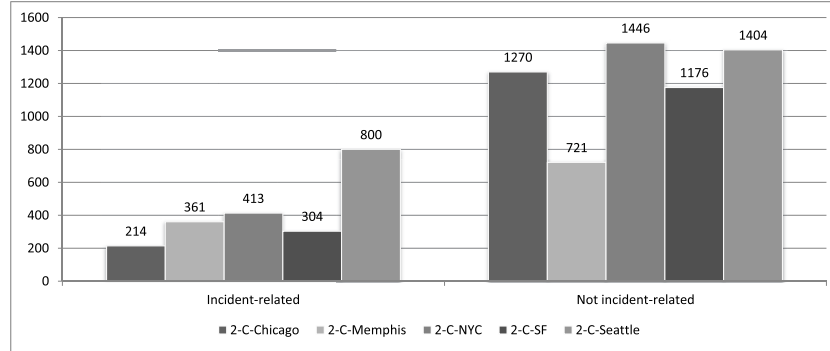


Figure 34.: Class distribution for 4-CLASSES and 2-CLASSES data sets.

Furthermore, we created ten additional data sets for evaluating the generalizability of our classifier. Two data sets have been created for each city, one for the two-class task and one for the four-class task. The class distributions of the two-class data sets are the following (see also Figure 22):

⁵¹ <http://www.crowdfunder.com/> [Accessed: 01.03.2014]

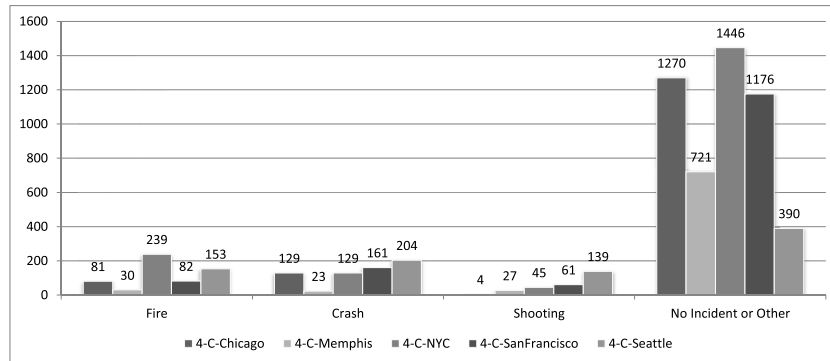
Table 22.: The two-class data sets for evaluating the semantic abstraction approach.



	Incident related	Not incident related	Total
2-C-Chicago	214	1,270	1,484
2-C-Memphis	361	721	1,082
2-C-NYC	413	1,446	1,859
2-C-SF	304	1,176	1,480
2-C-Seattle	800	1,404	2,204

The class distributions of the four-class data sets are shown in Figure 23. The distributions show that all data sets are imbalanced. Furthermore, the amount of tweets related to shootings is rather low, which might result in problems classifying this class.

Table 23.: The four-class data sets used for evaluating the semantic abstraction approach.



	Fire	Crash	Shooting	Not incident related or other	Total
4-C-Chicago	81	129	4	1,270	1,484
4-C-Memphis	30	23	27	721	801
4-C-NYC	239	129	45	1,446	1,859
4-C-SF	82	161	61	1,176	1,480
4-C-Seattle	153	204	139	390	886

Metrics

In the following evaluations, we provide metrics commonly used in information retrieval [245]:

$$\text{Accuracy (ACC)} = \frac{\text{Number of the correctly classified tweets}}{\text{Total number of tweets}} \quad (17)$$

$$\text{Precision (P)} = \frac{\text{Correctly classified positives}}{\text{Total predicted as positive}} \quad (18)$$

$$\text{Recall (R)} = \frac{\text{Correctly classified positives}}{\text{Total positives}} \quad (19)$$

$$\text{(balanced) F-measure (F)} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (20)$$

As we also deal with classifying multiple incident classes, we want to combine the individual measures for each class. Thus, we also provide the micro-averaged metrics [145]. This way, the class distribution is used for weighting the individual measures achieved for each class.

In the following, we present accuracy as well as F-measure as our main metrics. Furthermore, for comparing results, we always compare F-measures. This is because accuracy is not an appropriate measure for skewed data sets [145], which most social media data sets are.

Methodology

A standard approach for evaluating the performance of classifiers is to conduct cross validation [245]. In general, we used *k-fold cross validation* for evaluating our approaches. Thus, the data was partitioned into k equal-sized folds. For each run one fold was used for testing and the rest for training. This was repeated k times. The overall results were calculated based on the result of all k iterations. The evaluations were performed on stratified folds, which is important to keep the class distributions representative for each iteration.

Furthermore, for each iteration, we recalculated the TF-IDF scores based on the corpus of the current fold. Also, we optimized the cost parameter c whenever SVMs were used. Following [101], we tried exponentially growing sequences of c (e.g., $c = 2^5; 2^3; \dots; 2^{15}$). In the following results, we provide the best parameter settings for comparison.

7.5.2 Study 1: Incident Type Classification Based on Keywords

As mentioned before, we use a keyword-based prefiltering for selecting an initial set of tweets that is suitable for labeling. Thus, a first and simple approach for classifying incident-related tweets is to use these keywords for classification.

7.5.2.1 Study 1: Results

We applied these keywords used for keyword filtering in Section 3.2.2 to the 2-CLASSES as well as the 4-CLASSES data set and calculated the performance of this classification approach. For evaluating the four-class data set, the class with the largest number of keywords found in a tweet was chosen. If no decision could be made, the class "fire incident" was selected. The results for classifying each individual class as well as the weighted overall results are shown in Table 24 and Table 25.

We compared the results with a baseline, which is always predicting the majority class. The baseline achieves an accuracy of 51.45% and an F-measure of 34.96%. Compared with this baseline, using the keyword-based classification gives fair results with an accuracy of 50.15% and F-measure of 49.15%. Though the precision is quite high for the shooting class, it is rather low for predicting the other classes. A reason for this might be that shooting-related keywords are seldom used out of the context, whereas keywords such as "fire" frequently appear in daily chatter.

Table 24.: Classification results when keyword-based classification is applied on 4-CLASSES.

	Shooting	Fire	Crash	Not incident related	Micro Avg.
Precision	79.74%	44.34%	21.86%	52.88%	51.17%
Recall	55.38%	50.30%	28.80%	50.82%	48.09%
F-measure	65.36%	47.13%	24.85%	51.83%	49.15%

Conducting the same analysis on the two-class data set showed much worse performance compared with the baseline (Acc=64.67%, F=50.79%) with an accuracy of 48.17% and F-measure of 34.98%. The reason for this can be found in the nature of the data set that contains many false positives such as the tweet "Glad I realize you can't fight fire with fire....". This shows that a plain keyword-based approach needs more contextual information to predict accurately.

Table 25.: Classification results when keyword-based classification is applied on 2-CLASSES.

Precision	31.41%
Recall	39.45%
F-measure	34.98%

7.5.2.2 Study 1: Summary

Both results indicate that using a keyword-based classification is not sufficient and training a more accurate classifier is needed to get high classification performance.

7.5.3 Study 2: Incident Type Classification - Initial Feature Selection

In the following section, we present the study for finding the best feature set that enables high-quality classification of incident-related tweets. For this, we observed the influence of different feature combinations on precision and recall.

As outlined before, we focused on an NB and an SVM classifier as both showed to be the most valuable for text classification. For our evaluation, we used the Weka implementation of the multinomial NB model, which provided good results for other text classification tasks [151]. Further, we used the LibLinear [71] implementation of an SVM with linear kernel. We restricted our evaluation to a linear kernel, which has been shown as comparable to non-linear kernels for incident type classification tasks (see [38, 186]).

Though we conducted an intensive evaluation of multiple feature combinations, it is important to note that we did not evaluate all possible combinations, but only these that seem to provide better classification results. Furthermore, we were interested in finding the best feature combination that allows classifying both data sets. However, for future evaluations, one might also try to reduce the misclassification of certain classes.

7.5.3.1 Study 2: Results

In the following, we present the evaluation results. We started evaluating different n-gram combinations as well as weighting strategies as they provide a good baseline. Based on these results, we evaluated the value of adding each feature group exclusively as well as of applying replacement strategies. Based on this evaluation, we come up with an optimal feature combination using standard feature groups.

Based on these results, we evaluated our approach of semantic abstraction. First, we added each feature group exclusively to the best n-gram approach to show how these contribute. Second, we evaluated different combinations of the semantic abstraction feature groups in combination with the best feature groups identified before.

N-Grams and Weighting of N-Grams

We first evaluated how different n-gram combinations perform compared with a baseline approach predicting the majority class. The results are shown in Table 26 for the 4-CLASSES data set and Table 27 for the 2-CLASSES data set.

For the SVM, the results indicate that for the four-class data set, word-n-grams perform better compared with character-n-grams. Also, the results indicate that using smaller n for words and larger n for characters gives higher performance. For the two-class data set, the same effect could be verified for characters. However, for word-n-grams increasing n seems to be more beneficial. Furthermore, char-5-grams performed better compared with word-n-grams on this data set. In contrast, using the NB classifier, we found that using char-4-grams or char-5-grams outperforms word-n-grams on both data sets (4.48% on 4-CLASSES, 1.16% on 2-CLASSES).

In a second evaluation, we compared whether binary or TF-based weighting should be conducted. The analysis of the results for both weighting schemes shows that using SVM as a classifier, binary weighting outperforms TF-based weighting for both data sets. In contrast, using the NB classifier, for both data sets, TF-based weighting performs slightly better compared with binary weighting.

Regarding the overall performance, all approaches outperform the baseline approach. For SVMs, a maximum increase of 57.36% (4-CLASSES) and 38.94% (2-CLASSES) in F-measure is achieved. Also, the accuracy improves significantly. This shows that using even simple features such as word-n-grams gives high performance for this classification task. Also using the NB classifier, the baseline approach is outperformed by 53.94% (4-CLASSES) and 37.87% (2-CLASSES). Furthermore, the best approach for the NB classifier (88.90% and 88.70%) is outperformed by the SVM classifier (92.32% and 90.14%).

To finally decide which combination is used for further evaluations, we selected the top 15% results (see highlighted F-measures) and chose those approaches that perform best for both data sets. As a result of this analysis, we decided to use word-2-grams and word-3-grams for the SVM as well as char-5-grams for the NB approach. Furthermore, we use binary weighting for SVMs for each word-n-gram as it was beneficial compared with TF-based weighting. For NB, we conducted the following evaluations using the TF-based weighting.

Table 26.: Evaluation results for different n-gram combinations and weighting strategies on 4-CLASSES data set.

Classifier	c	Binary	Char-N-Gram	Word-N-Gram	Accuracy	F-measure
Majority Class					51.45%	34.96%
SVM	0.125	x		1	92.35	92.32%
	32768	x		2	92.15%	92.11%
	0.125	x		3	91.90%	91.84%
	8192	x		4	91.80%	91.75%
	128	x		5	91.60%	91.54%
	2	x	1		61.65%	56.46%
	0.03125	x	2		86.75%	86.68%
	0.03125	x	3		90.45%	90.42%
	0.03125	x	4		90.90%	90.86%
	0.125	x	5		91.50%	91.46%
	0.125			1	91.40%	91.38%
	0.5			2	91.85%	91.80%
	128			3	91.70%	91.64%
	512			4	91.50%	91.43%
	2048			5	91.40%	91.33%
	0.03125		1		68.70%	67.34%
	0.03125		2		85.15%	85.12%
	0.03125		3		89.60%	89.57%
	0.03125		4		90.40%	90.37%
	0.125		5		90.40%	90.37%
NB		x		1	84.45	84.42%
		x		2	79.25%	79.01%
		x		3	71.55%	70.53%
		x		4	66.00%	63.72%
		x		5	62.95%	59.55%
		x	1		57.10%	44.86%
		x	2		80.85%	80.55%
		x	3		86.40%	86.21%
		x	4		88.15%	87.99%
		x	5		88.90%	88.79%
				1	84.05%	84.01%
				2	79.95%	79.81%
				3	74.00%	73.38%
				4	68.30%	66.72%
				5	65.25%	62.78%
			1		65.00%	64.06%
			2		79.80%	79.65%
			3		86.70%	86.58%
			4		88.10%	87.95%
			5		89.05%	88.90%

Table 27.: Evaluation results for different n-gram combinations and weighting strategies on 2-CLASSES data set.

2 Classes						
Classifier	c	Binary	Char-N-Gram	Word-N-Gram	Accuracy	F-measure
Majority Class					64.67	50.79%
SVM	0.03125	x		1	88.98	88.83%
	0.5	x		2	89.99%	89.89%
	0.5	x		3	90.02%	89.90%
	0.5	x		4	89.84%	89.68%
	0.5	x		5	89.77%	89.61%
	0.03125	x	1		73.46%	72.27%
	0.03125	x	2		85.79%	85.66%
	0.03125	x	3		88.68%	88.62%
	0.03125	x	4		89.47%	89.42%
	0.03125	x	5		90.20%	90.14%
	0.03125			1	88.37%	88.22%
	0.125			2	89.59%	89.49%
	0.125			3	89.47%	89.32%
	0.03125			4	89.38%	89.17%
	2			5	89.38%	89.20%
	0.125		1		73.77%	72.23%
	0.03125		2		85.36%	85.25%
	0.03125		3		87.64%	87.60%
	0.03125		4		88.59%	88.52%
	0.03125		5		88.98%	88.92%
NB		x		1	87.37%	87.51%
		x		2	85.24%	85.50%
		x		3	80.37%	80.80%
		x		4	75.62%	76.07%
		x		5	71.09%	71.36%
		x	1		70.02%	62.34%
		x	2		82.84%	82.36%
		x	3		87.19%	86.93%
		x	4		88.47%	88.32%
		x	5		88.77%	88.70%
				1	87.28%	87.42%
				2	85.76%	86.00%
				3	81.86%	82.26%
				4	77.51%	77.97%
				5	73.68%	74.07%
			1		71.42%	70.03%
			2		80.31%	80.08%
			3		85.58%	85.41%
			4		88.07%	87.94%
			5		88.71%	88.66%

Replacement Strategies

In a second evaluation, we evaluated if applying the replacement strategies gives an increase of performance compared with not using them (see Table 28). The results show that slang replacement is indeed valuable for most cases. Solely replacing the URL is valuable for SVMs and four classes, but not for two classes. On the other hand, when NB is used as a classifier, the same approach is valuable for two classes, but not for four classes.

When combined with URL replacement, the results differ. For four classes, word-3-grams, and the SVM, the results increase slightly (+0.05%). Also for two classes and the NB classifier with char-5-grams, we get an increase (+0.06%). However, there is a slight drop for all other cases.

The results indicate that there is no clear improvement or decrease of performance, which is surprising as we would have expected an increase of performance. An analysis of the feature set showed that the overall amount of features is indeed decreased by 8%. One explanation for this phenomenon might be that the replaced tokens are not valuable for this classification task; thus, the results mostly remain stable. Also, outperforming the very good baseline is difficult. However, we decided to use slang, as well as URL replacement, as it helps to reduce the number of features and seems to be reasonable for social media data.

Syntactic Features

Adding the syntactic feature group consisting of the number of exclamation marks, questions marks, and the number of uppercase characters gives a slight improvement for most cases (+0.04% to +0.16%, see Table 29). Only if the NB classifier was applied on the 4-CLASSES, a small decrease of performance could be noticed. The results underline our initial assumption that syntactic differences help differentiate incident-related from unrelated tweets. As the results are mostly outperforming the baseline, we decided to include this feature group for further evaluations.

Emoticon Features

Adding the emoticon feature group decreases classification performance for four classes (see Section 30) up to 1.72%. A reason for this might be that emoticons do not help differentiate incident types. For two classes, the results are comparable, which could be an indication that emoticons have some effect on the classification task in this case. Nevertheless, with our current implementation of the emoticon feature group, we are not able to outperform the baseline. Based on these results, we decided not to use this feature group for further evaluations. For further evaluations, experimenting with different sentiment scores might be beneficial, at least for the two class case.

Table 28.: Evaluation results for replacement strategies before n-gram generation on 4-CLASSES and 2-CLASSES data sets.

4-CLASSES					
	Classifier	c	Accuracy	F-measure	
SVM	Baseline 2-Grams (bin)	32768	92.15%	92.11%	
	+ Slang	32768	91.95%	91.91%	-0.20%
	+ Slang + URL	8	91.95%	91.91%	-0.20%
	+ Url	32768	92.25%	92.21%	0.10%
	Baseline 3-Grams (bin)	0.125	91.90%	91.84%	
	+ Slang	0.125	92.00%	91.95%	0.10%
	+ Slang + URL	0.125	92.05%	92.00%	0.16%
	+ Url	8192	91.95%	91.90%	0.05%
NB	Baseline char-5-grams		89.05%	88.90%	
	+ Slang		89.10%	88.95%	0.05%
	+ Slang + URL		88.85%	88.75%	-0.15%
	+ Url		88.90%	88.80%	-0.10%
2-CLASSES					
		c	Accuracy	F-measure	
SVM	Baseline 2-Grams (bin)	0.5	89.99%	89.89%	
	+ Slang	0.5	90.17%	90.09%	0.20%
	+ Slang + URL	0.125	89.93%	89.82%	-0.07%
	+ Url	0.5	89.93%	89.85%	-0.04%
	Baseline 3-Grams (bin)	0.5	90.02%	89.90%	
	+ Slang	0.5	90.02%	89.90%	0.00%
	+ Slang + URL	0.125	89.99%	89.85%	-0.05%
	+ Url	0.125	89.93%	89.78%	-0.11%
NB	Baseline char-5-grams		88.71%	88.66%	
	+ Slang		88.80%	88.76%	0.09%
	+ Slang + URL		88.77%	88.75%	0.08%
	+ Url		88.74%	88.72%	0.06%

Table 29.: Evaluation results for syntactic features on 4-CLASSES and 2-CLASSES data sets.

4-CLASSES					
Classifier		c	Accuracy	F-measure	
SVM	Baseline 2-Grams (bin)	32768	92.15%	92.11%	
	+ Syntactic Features	0.5	92.25%	92.21%	0.10%
	Baseline 3-Grams (bin)	0.125	91.90%	91.84%	
	+ Syntactic Features	0.125	92.05%	92.00%	0.16%
NB	Baseline char-5-grams		89.05%	88.90%	
	+ Syntactic Features		88.95%	88.80%	-0.10%
2-CLASSES					
		c	Accuracy	F-measure	
SVM	Baseline 2-Grams (bin)	0.5	89.99%	89.89%	
	+ Syntactic Features	0.125	90.05%	89.93%	0.04%
	Baseline 3-Grams (bin)	0.5	90.02%	89.90%	
	+ Syntactic Features	0.5	90.17%	90.04%	0.15%
NB	Baseline char-5-grams		88.71%	88.66%	
	+ Syntactic Features		88.74%	88.69%	0.03%

Table 30.: Evaluation results for the emoticon features before n-gram generation on 4-CLASSES and 2-CLASSES data sets.

4-CLASSES					
Classifier		c	Accuracy	F-measure	
SVM	Baseline 2-Grams (bin)	32768	92.15%	92.11%	
	+Emo	0.5	92.05%	92.01%	-0.10%
	Baseline 3-Grams (bin)	0.125	91.90%	91.84%	
	+Emo	0.5	90.15%	90.12%	-1.72%
NB	Baseline char-5-grams		89.05%	88.90%	
	+Emo		89.05%	88.90%	0.00%
2-CLASSES					
		c	Accuracy	F-measure	
SVM	Baseline 2-Grams (bin)	0.5	89.99%	89.89%	
	+Emo	0.125	89.99%	89.87%	-0.02%
	Baseline 3-Grams (bin)	0.5	90.02%	89.90%	
	+Emo	0.125	89.99%	89.85%	-0.04%
NB	Baseline char-5-grams		88.71%	88.66%	
	+Emo		88.71%	88.66%	0.00%

TF-IDF

The result for the TF-IDF feature group is shown in Table 31. Adding TF-IDF features to the n-gram features shows that using the sum of TF-IDF scores gives comparable performance for NB (4-CLASSES: +0.05%, 2-CLASSES: +0%). However, using the SVM classifier, worse results are achieved with this feature. In contrast, using plain TF-IDF scores gives comparable or slightly better performance for both classifiers (4-CLASSES: +0.04% to +0.15%, 2-CLASSES: -0.04% to +0.07%). Using both feature groups, the results are worse for most cases.

A manual analysis of the sum of TF-IDF scores showed that high scores are mostly achieved for not incident-related tweets, whereas low scores are more prominent for incident-related tweets. Nevertheless, the sums do not seem to be a good differentiator for the different types of incidents, which might be an explanation for the bad results of the sum of TF-IDF scores. The weak increase of classification performance using the plain TF-IDF scores might be explained as these scores are not directly independent of the plain n-gram features; thus, the value for the overall classification task is not directly visible. However, TF-IDF scores provide good indicators for tokens that are representative for certain classes. Because of this and the slight increase of F-measure, we decided to use only TF-IDF scores for further evaluations.

7.5.3.2 Study 2: Summary

In this evaluation, we dealt with the problem of finding an optimal feature set for incident type classification of user-generated content. The evaluation of several n-gram combinations showed that for an SVM classifier, word-2-grams, and word-3-grams with binary weighting give the highest performance, and for an NB classifier, char-5-grams and TF weighting give the best results.

Though the replacement strategies did decrease the number of features, they only slightly increased classification performance. However, the reduction of the overall feature set is valuable for faster training; thus, we did not discard these strategies. Also, adding the syntactic feature group and TF-IDF scores showed to be beneficial. Emoticons and the sum of TF-IDF scores were not valuable for differentiating incident types; thus, they were discarded.

Table 31.: Evaluation results for TF-IDF scores on 4-CLASSES and 2-CLASSES data sets.

4-CLASSES					
	Classifier	c	Accuracy	F-measure	
SVM	Baseline 2-Grams (bin)	32768	92.15%	92.11%	
	+ TF-IDF scores	8192	92.30%	92.26%	0.15%
	+ TF-IDF sum	8	90.95%	90.96%	-1.15%
	+ TF-IDF both	8192	90.85%	90.86%	-1.25%
	Baseline 3-Grams (bin)	0.125	91.90%	91.84%	
	+ TF-IDF scores	128	91.95%	91.90%	0.05%
	+ TF-IDF sum	32	91.20%	91.19%	-0.65%
	+ TF-IDF both	512	91.10%	91.10%	-0.75%
NB	Baseline char-5-grams		89.05%	88.90%	
	+ TF-IDF scores		89.10%	88.93%	0.04%
	+ TF-IDF sum		89.10%	88.95%	0.05%
	+ TF-IDF both		89.05%	88.88%	-0.02%
2-CLASSES					
		c	Accuracy	F-measure	
SVM	Baseline 2-Grams (bin)	0.5	89.99%	89.89%	
	+ TF-IDF scores	0.5	90.05%	89.96%	0.07%
	+ TF-IDF sum	0.03125	89.26%	89.28%	-0.61%
	+ TF-IDF both	0.03125	89.23%	89.25%	-0.64%
	Baseline 3-Grams (bin)	0.5	90.02%	89.90%	
	+ TF-IDF scores	0.5	90.08%	89.95%	0.06%
	+ TF-IDF sum	0.125	89.41%	89.43%	-0.46%
	+ TF-IDF both	0.03125	89.50%	89.49%	-0.40%
NB	Baseline char-5-grams		88.71%	88.66%	
	+ TF-IDF scores		88.68%	88.62%	-0.04%
	+ TF-IDF sum		88.71%	88.66%	0.00%
	+ TF-IDF both		88.74%	88.68%	0.02%

7.5.4 Study 3: Incident Type Classification - Semantic Abstraction

In this evaluation, we aimed at showing that semantic abstraction is a valuable means for incident type classification. For the study, we focused on the two data sets 4-CLASSES and 2-CLASSES as both allow finding the best classifier and feature combination.

7.5.4.1 Results

In the following, we first present the results for evaluating each feature group generated using semantic abstraction compared with the plain n-gram approach. Second, we show how different combinations of the approaches for semantic abstraction perform. Third, we present that semantic abstraction is beneficial if the amount of training data is low.

Abstracting Temporal Mentions

Using the TEMP feature group gives comparable or slightly better performance for all cases (see Table 32). For four classes, the effects are rather low; however, using two classes shows that an increase of 0.06% for SVMs and 0.18% for NB can be achieved. The increasing performance can be explained with the reduced number of tokens. Furthermore, temporal expressions seem to be a good discriminator for this classification task.

Table 32.: Evaluation results for abstracting temporal expressions before n-gram generation on 4-CLASSES and 2-CLASSES data sets.

4-CLASSES					
Classifier		c	Accuracy	F-measure	
SVM	Baseline 2-Grams (bin)	32768	92.15%	92.11%	
	+Time	2	92.15%	92.11%	0.00%
	Baseline 3-Grams (bin)	0.125	91.90%	91.84%	
	+Time	2048	91.95%	91.90%	0.05%
NB	Baseline char-5-grams		89.05%	88.90%	
	+Time		89.00%	88.85%	-0.05%
2-CLASSES					
		c	Accuracy	F-measure	
SVM	Baseline 2-Grams (bin)	0.5	89.99%	89.89%	
	+Time	0.125	90.05%	89.94%	0.04%
	Baseline 3-Grams (bin)	0.5	90.02%	89.90%	
	+Time	0.5	90.08%	89.96%	0.06%
NB	Baseline char-5-grams		88.71%	88.66%	
	+Time		88.89%	88.85%	0.18%

Abstracting Location Mentions

Using the LOC feature group shows an increase for the two-class task using SVMs (see Table 33). Surprisingly, the results for the four-class task are slightly worse in the case an SVM is used. A reason for this could be that location mentions are not helpful for differentiating multiple incident types.

Using the NB classifier for the two as well as the four-class task showed worse results. The significant performance drop using the NB classifier might be a reason of using the char-5-grams, which are now less discriminative compared with not replacing the mentions.

Table 33.: Evaluation results for location mention abstraction before n-gram generation on 4-CLASSES and 2-CLASSES data sets.

4-CLASSES					
Classifier		c	Accuracy	F-measure	
SVM	Baseline 2-Grams (bin)	32768	92.15%	92.11%	
	+LOC	0.125	91.80%	91.78%	-0.33%
	Baseline 3-Grams (bin)	0.125	91.90%	91.84%	
	+LOC	0.5	91.65%	91.62%	-0.22%
NB	Baseline char-5-grams		89.05%	88.90%	
	+LOC		87.90%	87.81%	-1.09%
2-CLASSES					
		c	Accuracy	F-measure	
SVM	Baseline 2-Grams (bin)	0.5	89.99%	89.89%	
	+LOC	0.03125	90.47%	90.38%	0.49%
	Baseline 3-Grams (bin)	0.5	90.02%	89.90%	
	+LOC	0.125	90.47%	90.39%	0.50%
NB	Baseline char-5-grams		88.71%	88.66%	
	+LOC		86.61%	86.69%	-1.97%

Abstraction Based on Linked Open Data

We also compared the LOD feature group to the n-gram approach. The results show that in most cases, the performance drops (see Table 34). Also, separating the TYPES and CATEGORIES feature groups shows worse results. However, using this feature group for the two-class task and NB as a classifier gives an increase of F-measure. In the following evaluations we perform an in-depth evaluation of this feature group and show that it might not be helpful in differentiating incident types.

The results presented so far showed that using the single approaches of semantic abstraction in addition to the n-gram approach gives comparable and in some cases,

Table 34.: Evaluation results for the LOD feature groups before n-gram generation on 4-CLASSES and 2-CLASSES data sets.

4-CLASSES					
Classifier		c	Accuracy	F-measure	
SVM	Baseline 2-Grams (bin)	32768	92.15%	92.11%	
	+ALL	0.03125	90.40%	90.35%	-1.77%
	+TYPES	512	91.00%	90.96%	-1.15%
	+CATEGORIES	0.03125	91.05%	91.00%	-1.11%
	Baseline 3-Grams (bin)	0.125	91.90%	91.84%	
	+ALL	0.5	90.30%	90.24%	-1.60%
	+TYPES	8	90.55%	90.50%	-1.34%
	+CATEGORIES	0.125	91.00%	90.95%	-0.89%
NB	Baseline char-5-grams		89.05%	88.90%	
	+ALL		88.80%	88.68%	-0.21%
	+TYPES		89.10%	88.97%	0.08%
	+CATEGORIES		88.85%	88.70%	-0.20%
2-CLASSES					
		c	Accuracy	F-measure	
SVM	Baseline 2-Grams (bin)	0.5	89.99%	89.89%	
	+ALL	0.03125	89.38%	89.28%	-0.61%
	+TYPES	0.03125	89.81%	89.68%	-0.21%
	+CATEGORIES	0.125	89.62%	89.54%	-0.35%
	Baseline 3-Grams (bin)	0.5	90.02%	89.90%	
	+ALL	0.03125	89.47%	89.36%	-0.53%
	+TYPES	0.03125	89.68%	89.54%	-0.36%
	+CATEGORIES	0.125	89.65%	89.55%	-0.34%
NB	Baseline char-5-grams		88.71%	88.66%	
	+ALL		89.04%	89.01%	0.35%
	+TYPES		88.77%	88.73%	0.06%
	+CATEGORIES		88.95%	88.91%	0.24%

better performance for the TEMP and LOC feature groups. However, using LOD features did not show an increase of performance so far.

Semantic Abstraction

In the following, we compare different combinations of the LOD, LOC, and TEMP feature groups to find out if semantic abstraction is beneficial for classifying incident-related tweets. As a baseline we use the n-gram features after slang and URL replacement, TF-IDF scores, and syntactic features. For our evaluation, we also provide the *ALL* feature group, which is the combination of the LOD, LOC, and TEMP feature groups. In Table 35, the relative increase compared with the baseline is shown. The detailed results can be found in Chapter A.

For two classes, the results show that using all features in combination seems to be valuable for both n-gram cases and SVMs. Using only the LOC or TEMP feature group also gives better performance. Also, the combinations of LOC+TIME and LOC+LOD are beneficial. Using the LOD features in combination or separated shows worse performance. However, when NB is used as a classifier, the results are different as in this case, using the TYPES or the CATEGORIES feature groups gives a performance increase, whereas the other feature groups have a detrimental influence on the classification performance.

For four classes, the results indicate that only the TEMP feature group gives slightly better performance for SVMs; thus, the results from the two-classes study could not be verified in this case. However, when NB is used, the LOD feature group (in combination and also separated) gives an increase of performance. One reason for this might be that differentiating between two classes is easier and can be better supported with semantic abstraction compared with predicting four classes. In this case, abstracting temporal and location mentions might not help in differentiating incident types. This assumption is underlined by the positive results using NB and the LOD feature group, which provides more fine-grained information about the text at hand.

We showed that semantic abstraction can indeed be valuable for the two-class problem. Also, depending on the selection of the feature groups and the classifier, it can also be beneficial for the four-class problem. However, the current results provide only slight increases compared with the baseline approach, which is likely a reason of the large data sets used for training. However, as the following studies show, semantic abstraction is especially valuable when a generalized model needs to be trained.

Table 35.: Increase in F-measure for using semantic abstraction compared with a baseline comprising n-gram features after Slang and URL replacement, TF-IDF scores, and syntactic features.

Classes		N	+ALL	+LOC	+TIME	+LOD	+LOC +TIME
2	SVM	2	0.44%	0.62%	0.10%	-0.44%	0.73%
		3	0.19%	0.68%	0.22%	-0.26%	0.74%
	NB	5	-0.75%	-0.84%	-0.15%	-0.02%	-0.96%
4	SVM	2	-1.92%	-0.16%	0.09%	-1.67%	-0.30%
		3	-1.90%	-0.08%	0.05%	-1.66%	-0.23%
	NB	5	-0.66%	-0.76%	-0.43%	0.06%	-0.87%

Classes		N	+LOC +LOD	+TIME +LOD	+TYPES	+CAT.
2	SVM	2	0.34%	-0.47%	-0.31%	-0.43%
		3	0.43%	-0.41%	-0.16%	-0.28%
	NB	5	-0.90%	-0.03%	0.37%	0.20%
4	SVM	2	-1.78%	-1.92%	-1.26%	-1.11%
		3	-1.73%	-1.91%	-1.32%	-1.01%
	NB	5	-0.81%	0.06%	0.45%	0.29%

Evaluation of Semantic Abstraction for a Low Amount of Labeled Data

As assumed before, the value of semantic abstraction is not directly visible when applied to large data sets. Thus, we also evaluated semantic abstraction on smaller data sets. For evaluation, we used an "inverted" k -fold cross validation for which $(k - 1)$ folds are used for testing and one fold is used for training. We chose 60 (=1.67% of data set used for training), 50 (=2%), 40 (=2.5%), 30 (=3.33%), 20 (=5%), and 10 (=10%) folds. Furthermore, due to the large numbers of repetitions, we only evaluated an SVM classifier. We selected the same features that were used in the previous evaluation but used only word-3-grams.

In Figure 35, the learning curves for the 4-CLASSES data set are shown. The learning curve based on the CATEGORIES feature group performs best, even with this low amount of data. Furthermore, ALL, LOC+LOD, and TIME+LOD also provide better learning curves compared with the baseline. As can be seen, the learning curve for the baseline is faster increasing with a larger amount of data.

In Figure 36, the learning curves for the 2-CLASSES data set are shown. This time, the location abstraction performs best. Also, different combinations of LOC+TIME, LOC+LOD, and the complete semantic abstraction approach outperform the baseline approach with a low amount of training data.

The results indicate that semantic abstraction seems to be valuable when the amount of data used for training is low. For the 2-CLASSES data set, we get an increase of up to 5.54% in F-measure compared with the baseline, whereas the increase for the 4-CLASSES data set is up to 3.04%. This underlines the prior assumption that semantic abstraction is not as valuable with large training sets but provides better results with smaller sets.

7.5.4.2 Study 3: Summary

Our evaluations showed that the highest F-measure can be achieved using an SVM classifier. For both data sets, word-2-grams and word-3-grams in combination with slang and URL replacement, the syntactic feature group, TF-IDF scores, as well as different concepts for semantic abstraction achieve the highest performance.

For the 4-CLASSES data set we get an F-measure of 92.10% using word-2-grams and F=92.00% using word-3-grams in addition to the TIME feature group. For the 2-CLASSES data set, we get an F-measure of 90.70% for word-2-grams and 90.55% for word-3-grams in addition to the LOC+TIME feature groups. Both results show that (1) very good classification results can be achieved for incident type classification and that (2) semantic abstraction can be valuable for this task.

Furthermore, we showed that semantic abstraction provides much better results with smaller sets. In the following evaluation, we have a closer look at the performance of semantic abstraction with respect to creating a general classification model.

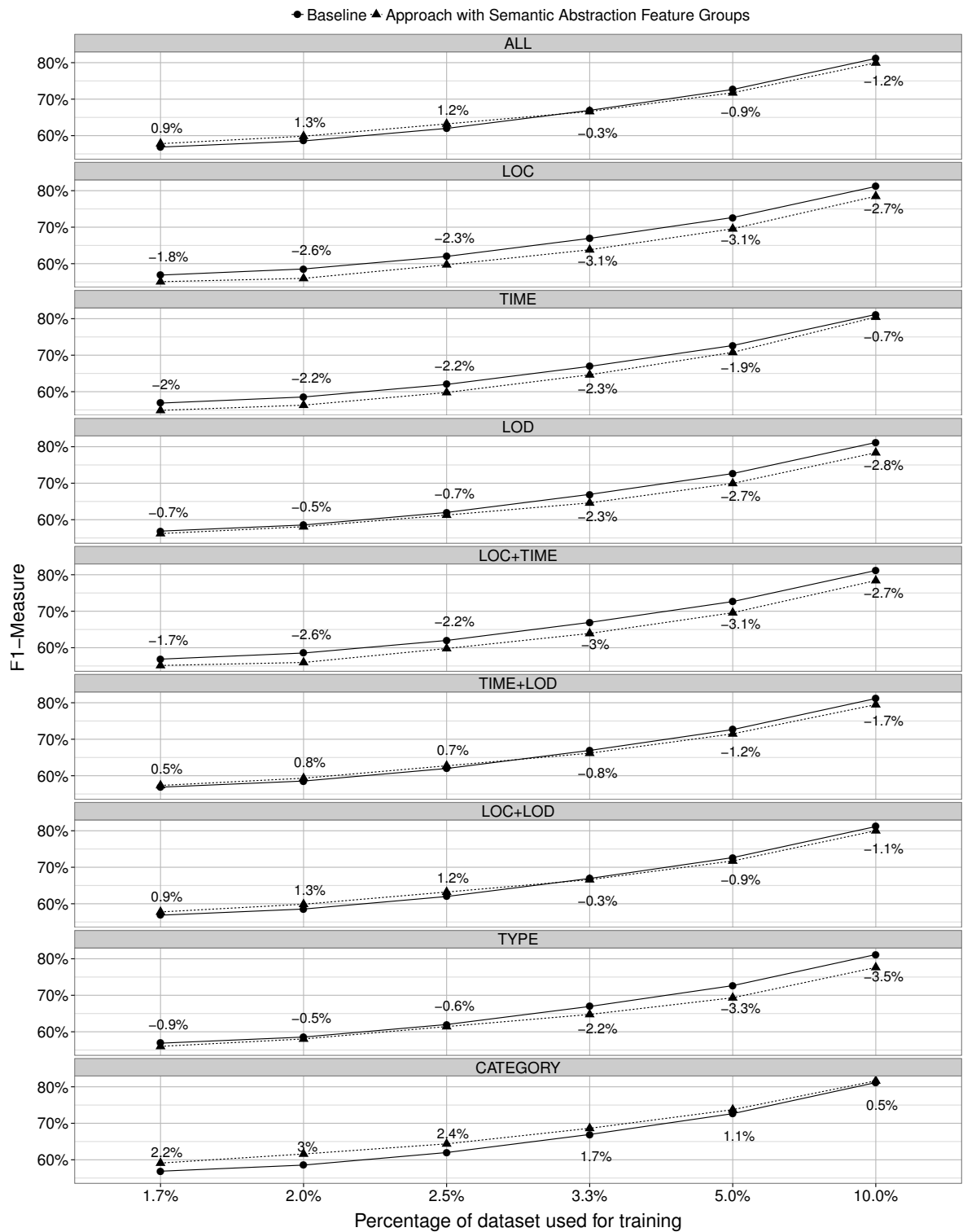


Figure 35.: Learning curves for different semantic abstraction approaches compared with the baseline on 4-CLASSES data set.

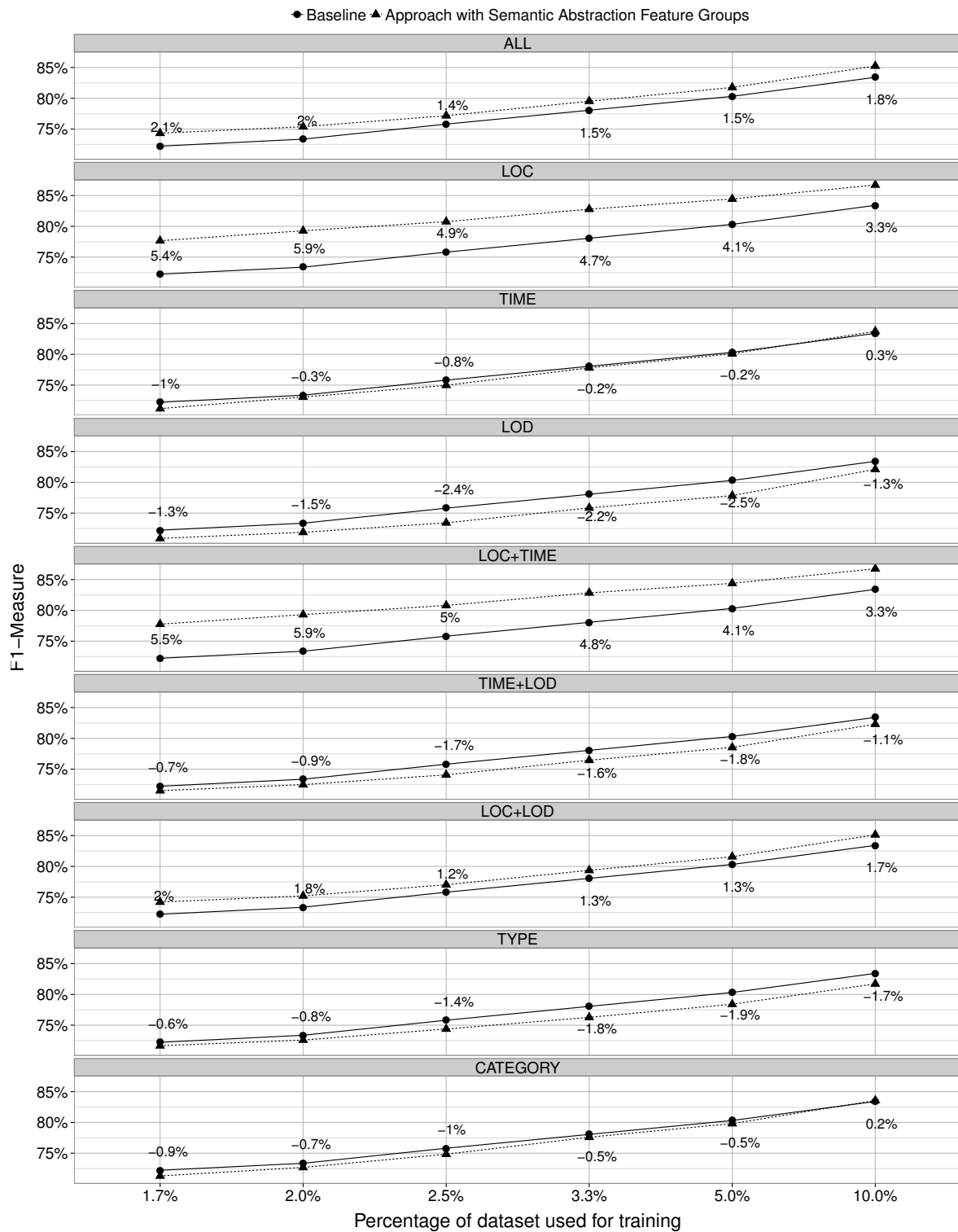


Figure 36.: Learning curves for different semantic abstraction approaches compared with the baseline on 2-CLASSES data set.

7.5.5 Study 4: Evaluation of Generalizability Using Semantic Abstraction

In this evaluation, we conducted two studies. The first study tackled the problem of how generalizable our approach is, especially when semantic abstraction is used. We show that semantic abstraction is a valuable means for creating a general model, although we also show that the results are ambiguous using the LOD feature group. Thus, we conducted a second study of the generalizability aspect, but this time, using a symbolic model, which allows interpreting the usage of single features. We will show that feature selection is especially needed for the LOD feature group.

7.5.5.1 Evaluation of Generalizability using Semantic Abstraction

As social media data varies considerably across different cities, we first present evaluation results that underline how generalizable our approach is. We conducted several studies using data from one city as training and data from a different city as a test set. We used the SVM classifier in combination with the best features evaluated before. Our expectation is that a model using features generated with semantic abstraction is more accurate on a different city compared with not using semantic abstraction.

In the following, we compare F-measures of different combinations of semantic abstraction feature groups with a baseline using the slang and URL replacement, the TF-IDF scores, as well as the syntactic feature group. Furthermore, we only use a SVM as classifier as this provided the best results in the prior evaluation. The results are shown as the increase or decrease in F-measure compared with the baseline.

Furthermore, we compared the baseline approach with the different combinations of applying semantic abstraction with respect to error rates. For this, we conducted McNemar's test [70] for evaluating the significance of the test results. We chose this test over a t-test because we cannot make any assumptions on the distribution of the classifier's performance measures [111]. To apply McNemar's test, for each instance, we checked how it was classified using the baseline approach and the respective approach using semantic abstraction. Based on this, we constructed the following contingency table:

Table 36.: Contingency table for conducting McNemar’s test.

NE_{00} : # of examples misclassified by both classifiers	NE_{01} : # of examples misclassified by classifier 1 but not by classifier 2
NE_{10} : # of examples misclassified by classifier 2 but not by classifier 1	NE_{11} : # of examples correctly classified by both classifiers

We assumed a difference in error rates, which resulted in the following hypothesis, which is tested for significance:

- H_1 : The error rates are different for two classifiers $NE_{01} \neq NE_{10}$
 $H_{1_0} : NE_{01} = NE_{10}$

In the following, we present the evaluation results for training on one data set and testing the remaining data sets.

Training on Chicago Data Sets

Training on the data sets that stem from Chicago and applying on all other data sets show that semantic abstraction yields better results compared with not using it. As a result of applying the significance test, we found that significantly better results could be achieved using semantic abstraction.

In the two-class case, using all feature groups for semantic abstraction shows significantly better performance compared with not using it. The highest increase of F-measure is achieved when the LOC feature group is used. Also, combining the LOC feature group with the TEMP and LOD feature groups shows high performance. In particular, when testing on 2-C-MEMPHIS, a gain of 13.50% could be achieved. Using only the CATEGORIES feature group shows worse results in most cases except when the model is applied on the 2-C-NYC data set (+3.58%). In contrast to this, using the TYPES feature group does increase the performance.

For the four-class case, the same results hold true for the LOC feature group and also the combinations of LOC and other feature groups. However, using all features groups for semantic abstraction gives only significantly better results for 4-C-MEMPHIS. For 4-C-SEATTLE, the results are even significantly worse, which is likely a reason of the class distribution in the 4-C-CHICAGO set. In general, when testing on 4-C-SEATTLE, the best results are achieved without semantic abstraction.

Training on Memphis Data Sets

Training on the Memphis data sets and applying on all other data sets show that semantic abstraction is also beneficial compared with not using it. As a result of applying the significance test, we found that by using semantic abstraction, significantly lower error rates could be achieved, especially when the LOC feature group is used.

Training and testing the two-class model show that semantic abstraction is beneficial when tested on 2-C-SEATTLE (10.53%). Also, the results on all data sets outperform the baseline whenever the LOC feature group is used. This also holds true in combination with the TEMP feature group. The results further show that using the TYPES and CATEGORIES feature groups gives better performance only when applied on the New York City (NYC) data set. In all other cases, these feature groups are not beneficial if used by themselves.

The same results hold true for the four-class case. Training on 4-C-MEMPHIS and testing on the other data sets shows that semantic abstraction, especially based on the LOC feature group, is beneficial, especially due to the fact that 4-C-MEMPHIS has a very low number of incident-related tweets; thus, training a generalized model is much more complicating. Using the LOD or TEMP feature group shows worse results, whereas the combination of both can give better results. Using the TYPES or CATEGORIES feature groups does not increase the performance in most cases.

Training on New York City Data Sets

Training on the NYC data sets and applying on all other data sets also show that semantic abstraction is also beneficial compared with not using it. As a result of applying the significance test, we found that by using semantic abstraction, significantly lower error rates could be achieved, especially when the LOC feature group is used. Also in this case, the TEMP and LOD feature groups show significantly lower error rates in some cases.

When testing the two-class model, semantic abstraction is useful except when tested on 2-C-CHICAGO. Also in this combination, the LOC feature group gives the highest performance, which also holds true for combinations with other feature groups. However, the TEMP and LOD feature groups are also beneficial when the set is applied on 2-C-SEATTLE. Nevertheless, the LOD feature group also leads to lower performance when the model is applied on 2-C-CHICAGO. The TYPES and CATEGORIES are useful in some cases but can also lead to lower performance.

The same results could be verified for the four-class data sets for which semantic abstraction is also useful. The LOC feature group gives the highest performance, also in combinations, whereas the TEMP and LOD feature groups can be beneficial but also detrimental for the performance.

Table 37.: Evaluation of semantic abstraction when trained on Chicago and Memphis data sets. (* significant difference in error rates with $p < 0.05$, ** with $p < 0.01$)

		Baseline	+ALL	+LOC	+TIME	+LOD	+LOC+TIME	+LOC+LOD	+TIME+LOD	+TYPES	+CAT.
Memphis	2	65.33%	4.92%	** 13.18%	** 0.25%	1.42%	* 13.5%	** 1.87%	* 5.53%	** 1.36%	* -0.73%
	3	63.31%	6.25%	** 12.51%	** 0.30%	2.85%	** 12.92%	** 3.00%	** 5.71%	** 1.52%	-0.03%
	2	80.50%	1.29%	1.11%	0.33%	2.51%	* 1.31%	2.74%	** 1.82%	1.46%	1.78%
	3	77.35%	4.57%	** 1.99%	1.23%	4.63%	** 2.12%	5.07%	** 3.74%	** 3.48%	** 3.58%
	2	83.10%	0.20%	2.07%	0.75%	-1.40%	** 2.35%	* -1.44%	** 0.42%	-0.97%	* -0.94%
SF	3	80.61%	2.48%	2.86%	** 0.67%	0.65%	2.85%	** 0.85%	2.17%	0.76%	0.71%
	2	66.62%	7.09%	** 9.52%	** 1.08%	* 0.86%	9.14%	** 1.96%	6.71%	** 2.47%	-1.09%
Seattle	3	64.27%	9.49%	** 10.87%	** 1.20%	* 3.27%	* 11.74%	** 3.45%	7.06%	** 4.70%	** -0.29%

		Baseline	+ALL	+LOC	+TIME	+LOD	+LOC+TIME	+LOC+LOD	+TIME+LOD	+TYPES	+CAT.
Memphis	2	86.60%	0.73%	1.48%	0.11%	-0.51%	* 1.26%	-0.35%	* 0.64%	-0.58%	-0.14%
	3	86.10%	1.27%	2.26%	** 0.21%	0.07%	1.92%	* -0.02%	1.30%	-0.17%	0.04%
	2	79.70%	2.45%	1.27%	* 0.36%	1.85%	** 1.86%	* 1.63%	* 1.45%	* 0.88%	0.89%
NYC	3	77.30%	2.99%	** 1.34%	** 0.11%	** 3.19%	** 5.33%	** 3.62%	** 3.14%	** 1.77%	** 2.01%
	2	81.40%	-0.44%	2.02%	* 0.64%	-1.55%	2.59%	** -1.32%	-0.71%	-1.39%	-1.12%
	3	80.60%	-0.81%	0.81%	** -1.31%	-1.38%	1.34%	** -1.50%	-0.51%	-2.33%	-1.03%
Seattle	2	44.10%	5.33%	6.92%	** 1.63%	1.02%	7.09%	** 1.17%	4.50%	* 0.96%	-1.73%
	3	52.8%	-5.22%	** 1.66%	-10.66%	** -8.75%	6.16%	** -8.20%	** -4.75%	** -9.48%	** -10.93%

		Baseline	+ALL	+LOC	+TIME	+LOD	+LOC+TIME	+LOC+LOD	+TIME+LOD	+TYPES	+CAT.
Chicago	2	87.70%	0.34%	1.76%	-0.08%	0.12%	1.26%	-0.35%	* 0.64%	-0.58%	-0.14%
	3	86.90%	0.77%	2.27%	-0.06%	0.68%	1.92%	* -0.02%	1.3%	-0.17%	0.04%
NYC	2	75.00%	2.61%	3.47%	0.07%	1.50%	1.86%	** 1.63%	* 1.45%	* 0.88%	0.89%
	3	74.1%	3.31%	3.02%	0.34%	2.36%	5.33%	** 3.62%	** 3.14%	** 1.77%	** 2.01%
SF	2	86.60%	-1.67%	** 1.23%	0.44%	-1.62%	2.59%	** -1.32%	-0.71%	-1.39%	-1.12%
	3	85.70%	-0.60%	* 1.53%	-0.09%	-1.00%	1.34%	** -1.50%	-0.51%	-2.33%	-1.03%
Seattle	2	67.70%	10.53%	** 14.58%	** 0.21%	3.45%	** 7.09%	** 1.17%	4.50%	* 0.96%	-1.73%
	3	66.7%	12.05%	** 15.47%	** -0.21%	4.05%	** 6.16%	** -8.20%	** -4.75%	** -9.48%	** -10.93%

		Baseline	+ALL	+LOC	+TIME	+LOD	+LOC+TIME	+LOC+LOD	+TIME+LOD	+TYPES	+CAT.
Chicago	2	80.80%	1.55%	3.23%	** -0.36%	0.08%	2.72%	** 0.33%	1.37%	0.35%	0.46%
	3	80.50%	2.14%	* 2.19%	** -0.25%	0.21%	1.38%	* 0.98%	2.25%	* -0.42%	0.31%
NYC	2	69.80%	0.75%	1.54%	-0.31%	0.02%	1.55%	** 0.20%	2.13%	0.31%	* 0.85%
	3	69.10%	1.22%	3.04%	0.18%	0.81%	2.34%	0.50%	1.27%	0.27%	1.47%
SF	2	76.40%	-0.63%	** 2.56%	** -1.43%	** -1.69%	** 1.18%	-1.66%	** -1.30%	* -0.10%	-0.73%
	3	75.6%	-0.72%	2.85%	-1.02%	* -1.26%	* 0.07%	-1.15%	** -0.53%	-2.16%	** -0.37%
Seattle	2	54.50%	6.14%	16.71%	** -2.33%	** -1.28%	10.55%	** -0.50%	6.38%	** -0.29%	-0.26%
	3	51.30%	6.54%	** 13.12%	** -1.47%	0.07%	6.30%	** 0.36%	6.93%	** -1.60%	** -3.11%

(a) 2-C-CHICAGO

(b) 4-C-CHICAGO

(c) 2-C-MEMPHIS

(d) 4-C-MEMPHIS

Training on San Francisco Data Sets

If the San Francisco data sets are used for training, semantic abstraction showed to be not useful as it was for the other data sets. Nevertheless, for all test data sets, there are feature combinations that lead to significantly lower error rates compared with the baseline.

For the two-class model, all feature groups for semantic abstraction are only useful when applied on 2-C-NYC. However, using only the LOC feature group can give an increase in performance, except for the 2-C-SEATTLE data set. For this data set, the combination of LOC and TEMP feature groups gives the best results. Nevertheless, using the TEMP and LOD feature groups shows worse performance in most cases. Using TYPES and CATEGORIES is only beneficial when applied on the NYC data set.

The four-class model shows the same effects on data set 4-C-NYC. However, also for the 4-C-SEATTLE data set, much better results are achieved when all feature groups are used. Using the LOC feature group is only beneficial for 4-C-CHICAGO and 4-C-NYC. In this case, the TEMP feature group is much more valuable compared with the two-class case and leads to improved performance. TYPES and CATEGORIES are once again only valuable when the model is applied on the NYC data set.

Training on Seattle Data Sets

Using semantic abstraction on the Seattle data sets also shows that it is beneficial to not use it. Furthermore, we could find a significant difference of error rates for the two-class experiments except when applied on 2-C-SF. For the four-class model, we could only verify significantly better error rates when applied on 4-C-CHICAGO and 4-C-NYC.

Training on 2-C-SEATTLE and applying on the other two-class data sets show that semantic abstraction is indeed valuable as it increases F-measure. Only when applied on the 2-C-SF data set are the results slightly worse. Furthermore, different combinations of the LOC and TEMP feature groups also show good performance. However, using the LOD feature group in combination or as a single feature results in decreased performance.

When trained and tested on the four-class data set, semantic abstraction does not show to be as valuable compared with all other experiments conducted before. However, using only the TEMP and LOC feature groups shows comparable performance in most cases. Testing on 4-C-MEMPHIS and 4-C-SF shows only a slight improvement over the baseline.

Table 39.: Evaluation of semantic abstraction when trained on New York City and San Francisco data sets. (* significant difference in error rates with $p < 0.05$, ** with $p < 0.01$)

		+ALL	+LOC	+TIME	+LOD	+LOC+TIME	+LOC+LOD	+TIME+LOD	+TYPES	+CAT.
(a) 2-C-NYC	N	Baseline								
	2	87.89%	-1.12%	-0.64%	-2.16%	0.54%	-2.11%	-1.31%	-0.20%	-2.58%
	3	85.73%	-0.21%	-0.45%	-1.13%	1.79%	-0.97%	0.15%	0.32%	-1.27%
	2	65.00%	6.59%	-0.38%	4.01%	9.48%	3.81%	6.47%	4.83%	-0.31%
	3	59.51%	11.25%	0.17%	6.48%	11.12%	6.82%	11.50%	6.95%	3.65%
	2	81.74%	2.37%	0.56%	1.59%	3.74%	1.56%	2.34%	0.71%	1.21%
	3	80.23%	2.48%	-0.33%	2.24%	2.94%	2.54%	2.82%	1.37%	1.55%
	2	65.18%	7.58%	2.59%	3.18%	12.64%	3.95%	6.38%	4.35%	-0.08%
	3	60.10%	9.90%	2.83%	6.01%	15.61%	6.30%	10.02%	5.88%	3.19%
	N	Baseline								
	2	85.97%	0.94%	-0.54%	0.72%	1.40%	-0.62%	0.74%	0.54%	-1.91%
	3	84.37%	1.61%	-0.74%	-0.09%	2.22%	-0.09%	1.56%	0.90%	0.47%
(b) 4-C-NYC	2	86.26%	0.18%	0.01%	-0.28%	1.96%	0.04%	-0.05%	-0.22%	-0.87%
	3	85.67%	0.70%	0.07%	-0.11%	2.25%	0.03%	0.70%	0.14%	-0.66%
	2	79.77%	1.00%	0.41%	0.27%	2.35%	0.21%	0.98%	-1.00%	1.10%
	3	76.94%	3.13%	-0.55%	2.53%	2.82%	2.43%	3.24%	0.46%	3.86%
	2	42.05%	10.56%	3.42%	3.20%	14.50%	4.60%	10.25%	3.31%	-1.19%
	3	36.19%	14.39%	3.99%	6.61%	18.56%	5.97%	13.91%	5.40%	3.93%
	N	Baseline								
	2	89.30%	-1.32%	-0.09%	-1.75%	0.24%	-1.87%	-1.68%	-1.84%	-1.10%
	3	87.66%	-0.48%	0.47%	0.16%	1.32%	-0.68%	-0.63%	-0.94%	-0.21%
	2	82.18%	-8.07%	-5.14%	-10.19%	1.75%	-10.39%	-7.90%	-10.60%	-7.62%
	3	76.09%	-3.24%	-0.81%	-3.05%	4.55%	-5.46%	-3.24%	-6.19%	-2.38%
	2	78.95%	2.03%	0.34%	1.81%	0.73%	1.78%	1.89%	2.11%	1.07%
	3	75.18%	4.13%	0.61%	5.32%	1.51%	3.86%	4.00%	3.95%	2.98%
(c) 2-C-SF	2	74.85%	-4.88%	0.27%	-4.67%	0.31%	-4.86%	-3.27%	-2.33%	0.38%
	3	70.80%	0.18%	0.20%	0.51%	2.26%	-1.33%	-0.08%	-0.13%	1.96%
	N	Baseline								
	2	89.22%	-1.61%	1.89%	-2.46%	0.84%	-2.49%	-1.86%	-1.72%	-1.65%
	3	87.38%	-0.43%	0.27%	-0.86%	1.3%	-1.16%	-0.25%	-1.20%	-0.83%
	2	89.17%	-0.29%	-0.32%	-0.29%	0.00%	-0.35%	0.23%	-1.12%	0.23%
	3	89.25%	0.24%	-0.31%	-0.47%	-0.92%	-0.51%	-0.29%	-1.39%	0.42%
	2	78.18%	2.13%	7.35%	2.14%	1.44%	2.56%	2.30%	0.83%	1.88%
	3	73.70%	6.11%	0.21%	5.33%	1.65%	5.11%	6.38%	3.36%	3.88%
	2	56.22%	-0.44%	8.51%	0.32%	-1.79%	0.02%	0.72%	-4.19%	2.68%
	3	45.65%	7.57%	1.55%	6.04%	2.48%	5.42%	8.28%	5.99%	9.15%
	N	Baseline								
	2	89.22%	-1.61%	1.89%	-2.46%	0.84%	-2.49%	-1.86%	-1.72%	-1.65%
	3	87.38%	-0.43%	0.27%	-0.86%	1.3%	-1.16%	-0.25%	-1.20%	-0.83%
(d) 4-C-SF	2	89.17%	-0.29%	-0.32%	-0.29%	0.00%	-0.35%	0.23%	-1.12%	0.23%
	3	89.25%	0.24%	-0.31%	-0.47%	-0.92%	-0.51%	-0.29%	-1.39%	0.42%
	2	78.18%	2.13%	7.35%	2.14%	1.44%	2.56%	2.30%	0.83%	1.88%
	3	73.70%	6.11%	0.21%	5.33%	1.65%	5.11%	6.38%	3.36%	3.88%
	2	56.22%	-0.44%	8.51%	0.32%	-1.79%	0.02%	0.72%	-4.19%	2.68%
	3	45.65%	7.57%	1.55%	6.04%	2.48%	5.42%	8.28%	5.99%	9.15%

(a) 2-C-NYC

(b) 4-C-NYC

(c) 2-C-SF

(d) 4-C-SF

Table 41.: Evaluation of semantic abstraction when trained on Seattle data sets. (* significant difference in error rates with $p < 0.05$, ** with $p < 0.01$)

	N	Baseline	+ALL	+LOC	+TIME	+LOD	+LOC+TIME	+LOC+LOD	+TIME+LOD	+TYPES	+CAT.
Chicago	2	89.14%	1.84%	*	-0.31%	0.87%	0.19%	0.87%	1.54%	*	1.12%
	3	88.92%	1.98%	*	-0.46%	0.94%	-0.11%	0.91%	1.88%	*	0.55%
Memphis	2	83.20%	3.10%	**	0.44%	-1.47%	4.54%	** -1.74%	1.34%	-2.32%	** -0.07%
	3	83.74%	2.17%	*	0.37%	-2.00%	4.09%	** -1.36%	1.88%	-3.18%	* -1.29%
NYC	2	83.78%	-0.31%	*	-0.42%	0.40%	-2.82%	** 0.14%	0.11%	1.12%	*
	3	82.08%	1.25%	*	0.29%	1.56%	-1.74%	** 1.60%	2.16%	1.96%	** 0.96%
SF	2	87.31%	-0.41%		0.17%	-0.22%	0.11%	-0.48%	0.24%	-0.85%	0.54%
	3	87.24%	-0.43%		-0.13%	-0.30%	-0.68%	-0.16%	0.02%	-0.65%	0.46%

(a) 2-C-Seattle

	N	Baseline	+ALL	+LOC	+TIME	+LOD	+LOC+TIME	+LOC+LOD	+TIME+LOD	+TYPES	+CAT.
Chicago	2	88.43%	-0.55%	0.61%	0.27%	-1.79%	0.60%	-1.99%	-0.93%	-1.06%	** -3.64%
	3	86.32%	2.12%	2.59%	2.12%	-0.05%	2.20%	-0.12%	2.44%	1.52%	** -2.10%
Memphis	2	93.3%	-0.27%	-0.02%	0.13%	-2.51%	0.18%	-0.79%	-1.01%	-1.37%	** -1.89%
	3	93.54%	-0.76%	-0.06%	0.14%	-2.26%	0.04%	-1.73%	-0.68%	-0.88%	** -2.09%
NYC	2	84.00%	-0.23%	0.46%	1.27%	-1.04%	0.63%	-1.24%	0.15%	-1.72%	-0.75%
	3	82.92%	0.55%	0.63%	2.06%	-0.28%	2.49%	-0.41%	0.50%	-0.12%	-0.70%
SF	2	85.94%	-1.80%	0.09%	-0.04%	-1.84%	0.16%	-1.95%	-2.17%	-2.22%	* -1.40%
	3	85.21%	-1.87%	0.65%	1.31%	-0.93%	0.55%	-0.98%	-1.73%	0.04%	** -0.59%

(b) 4-C-Seattle

In this evaluation, we showed the value of semantic abstraction on five diverse data sets. The evaluation results indicate that by using semantic abstraction, a better performance can be achieved using either all feature groups or at least one single feature group in addition to the baseline approach. In particular, using the LOC feature group seems to be valuable for creating a generalized model. Semantic abstraction seems to be most valuable for the two-class case as it helps to differentiate incident-related tweets from not incident-related tweets. This differentiation is much more difficult for several incident classes, which was also shown in the previous evaluation.

For the TEMP and LOD feature groups, the results are heterogeneous as in some cases a significant performance increase is achieved and in other cases is not achieved. Surprisingly, the LOD feature group did not perform well in all cases. We found that using the TYPES feature group is more beneficial compared with using the CATEGORIES feature group. This is why we perform an in-depth evaluation of semantic abstraction, especially on this feature group, in the following section.

In this study, we showed that semantic abstraction is especially valuable when it comes to training a general model that is trained on one city and applied on a different one. We could further prove the significance of our results with respect to the error rates.

7.5.5.2 Evaluation of Semantic Abstraction Using a Rule-Based Classifier

Though we could show an improved classification performance using semantic abstraction, it was quite surprising that the LOD feature group performed worse in many cases. Compared with using the LOC and TEMP feature groups, we expected to see a positive influence for the generalization problem. Thus, in the following section, we present the results of an in-depth evaluation of semantic abstraction using a symbolic model that allows the interpretation of the usage of single features.

The different features are combined and evaluated using the machine learning library Weka and the Ripper rule learner (JRip) algorithm [245]. We decided to use a rule learning algorithm to be able to interpret the resulting models. As our primary interest is to evaluate the importance of semantic abstraction, we are not interested in finding the model that yields the highest F-measure. Albeit statistical models might have a better performance than symbolic ones, they are not interpretable and therefore not applicable for our purpose.

The advantage is that we can easily identify what rule was used to classify the example at hand, what conditions this rule consists of, and, consequently, what features were used. Then we can measure how often abstract features were used instead of the regular ones (e.g., bag of words). Also, we are interested in finding rule sets that contain many rules. Consequently, we did experiments with the unpruned version

of JRip as the pruned one often ends up in rule sets with very few and very general rules for which it can be hard to see any direct influence of the semantic abstraction.

Our evaluation is focused on the 2-C-MEMPHIS and 2-C-SEATTLE data sets as they provide us with sufficient information to understand the usage of the LOD feature group. To evaluate the learned rule sets we used one run of a tenfold cross validation whenever no test set was present (i.e., the cases when we evaluate on tweets of a single city). In all other cases, we use one city as a training set and the other city as a test set.

Furthermore, as we did not conduct our feature selection for the JRip algorithm, we use word-3-grams, TF-IDF scores and the sum of TF-IDF scores, syntactic features, as well as slang and URL replacement. Same as before, all combinations of the feature groups for semantic abstraction are evaluated.

In the following, we first give an overview of statistics of the data sets. Next, we show our results when the classifier is evaluated on one city, and third, we present results when training and testing on data from different cities. Fourth, we analyze first approaches for optimizing the usage of LOD features for the two-city classification problem.

Data Set Statistics

As we aimed at using two heterogeneous data sets from two cities, we analyzed how similar they are. Table 43 shows the overall number of unique tokens before and after preprocessing. The results indicate that after preprocessing, 28% of all commonly shared tokens are present in the Seattle data set and 48% in the Memphis data set. This shows that there are indeed huge differences between the tokens of both cities. This emphasizes the initial hypothesis that using plain n-grams is not sufficient for achieving high classification results on such diverse data sets. Furthermore, the results show the importance of applying the preprocessing of tweets to get a common base of tokens for feature generation.

Table 43.: Number of tokens both data sets have in common.

	Unprocessed	Processed
2-C-SEATTLE	10339	3606
2-C-MEMPHIS	5657	2070
$2\text{-C-SEATTLE} \cap 2\text{-C-MEMPHIS}$	1993	1007

In Table 44, the number of tweets for which location and temporal expressions, as well as LOD features, could be extracted is shown. The results indicate that location mentions and LOD features could be extracted for about 50% of all tweets in both data sets. Furthermore, temporal expressions could be identified in only 20%. The table also shows that for more than 37% of all tweets, location mentions

and LOD features could be extracted in one tweet. This is likely to be a result that location mentions are also linked to URIs using Spotlight. Further, taking temporal expressions into account reduces the number significantly.

Table 44.: Number of tweets containing location mentions and temporal expressions as well as LOD types and categories.

	2-C-SEATTLE	2-C-MEMPHIS
LOC	1295 (58.76)%	522 (48.24%)
TEMP	403 (18.28%)	265 (24.49%)
Types	1269 (57.58%)	566 (52.31%)
Categories	1222 (55.44%)	548 (50.65%)
ALL	160 (7.26%)	106 (9.80%)
LOC + TEMP	256 (11.62%)	140 (12.94%)
LOC + LOD	873 (39.61%)	409 (37.80%)
TEMP + LOD	254 (11.52%)	161 (14.88%)

Furthermore, we analyzed the number of distinct types and categories that could be extracted for both data sets (see Table 45). Comparing the LOD features for both cities shows that 880 types and 1,553 categories are shared by both data sets. On the one hand, this means that LOD features are indeed helpful; on the other hand, a feature selection seems to be necessary.

Table 45.: Number of distinct types and categories extracted for both data sets.

	2-C-SEATTLE	2-C-MEMPHIS
Distinct Types	3037	1553
Distinct Categories	4812	2042
2-C-SEATTLE \cap 2-C-MEMPHIS Types	880	880
2-C-SEATTLE \cap 2-C-MEMPHIS Categories	1553	1553

We also analyzed the five most representative LOD features for both classes in both data sets. The representativeness was calculated based on the number of incident-related and not incident-related tweets containing a certain LOD feature. On the one hand, the results in Table 46 indicate that mostly types and categories related to location mentions are relevant for incident-related tweets. As shown, both data sets have many of these LOD features in common. On the other hand, a variety of different LOD features are present for tweets not related to incidents. In this case, both data sets have a very limited number of LOD features in common.

Table 46.: The most representative LOD features for incident-related and not incident-related tweets in each data set.

2-C-SEATTLE	2-C-MEMPHIS
Incident-Related Tweets	
../ontology/ArchitecturalStructure	../ontology/Place
../ontology/Infrastructure	../ontology/Infrastructure
../ontology/RouteOfTransportation	../ontology/ArchitecturalStructure
../ontology/Road	../ontology/Road
../resource/Category:Interstate_5	../ontology/RouteOfTransportation
...	...
../class/yago/YagoLegalActorGeo	../class/yago/Conveyance103100490
../class/yago/YagoPermanentlyLocatedEntity	../ontology/MeanOfTransportation
../class/yago/YagoLegalActor	../ontology/Automobile
../ontology/Agent	../resource/Category:Living_people
../class/yago/Abstraction100002137	../class/yago/Instrumentality103575240
Not incident-Related Tweets	

Using Tweets From One City Only

In the first experiment, we wanted to see how important semantic abstraction is when we use data from one city only. As we implicitly follow two goals (i.e., to generalize to unseen data from the same city and to generalize to a completely different city), we start by giving results for one city. In Table 47, the results of applying different feature combinations on both data sets are shown. Furthermore, we provide a baseline using the majority class.

The results for 2-C-MEMPHIS indicate that using all features results in the best classification performance ($F = 85.80\%$). Compared with not using semantic abstraction ($F = 83.66\%$), we get an increase of 2.14%. However, the results on this data set also show that using temporal expressions and LOD categories decreases the classification results.

For the 2-C-SEATTLE data set, we get an increase of 1.94% by using semantic abstraction. In this case, except for the combination of LOD and temporal expressions, all feature combinations improve the classification results. It seems that semantic abstraction is indeed a valuable means for classification of data sets derived from one city and that a combination of all features works best.

Generalizing From One City to Another One

The classification results for training a classifier on one city and applying it on the other city are shown in Table 48. They indicate that using semantic abstraction outperforms the simple approach without semantic abstraction by 8.24% and 7.29%,

Table 47.: F-measures for training and testing on one data set using 10f-CV.

	2-C-MEMPHIS	2-C-SEATTLE
ALL	85.80%	81.17%
LOC + TEMP	85.65%	80.52%
LOD + TEMP	85.23%	78.75%
LOD + LOC	85.26%	81.32%
LOD	85.42%	79.40%
TYPES	85.33%	79.40%
CATEGORIES	83.48%	79.19%
TEMP	82.45%	79.40%
LOC	84.60%	79.47%
No SemAbs.	83.66%	79.23%
Majority Class	53.30%	49.58%

respectively. However, training a model on 2-C-SEATTLE and applying it on 2-C-MEMPHIS tweets shows that LOC + TEMP features provide the best results. TEMP and LOC are both valuable feature groups for the classification problem compared with not using semantic abstraction. However, the results also show that using just LOD features results in a significant drop of classification performance, although for the 2-C-MEMPHIS to 2-C-SEATTLE evaluation, using LOD features in combination with the other feature groups yielded the best results. This is likely to be the case because the combination of all features allows the finer differentiation of LOD features even if they do not work well in isolation.

Table 48.: F-measures for training on one city and testing on a different city.

	2-C-MEMPHIS to 2-C-SEATTLE	2-C-SEATTLE to 2-C-MEMPHIS
ALL	81.40% (+8.24%)	79.07% (+7.32%)
LOC + TEMP	80.43% (+7.27%)	80.58% (+8.82%)
LOD + TEMP	55.64% (-17.52%)	71.29% (-0.46%)
LOD + LOC	69.39% (-3.78%)	74.89% (+3.14%)
LOD	64.84% (-8.32%)	64.00% (-7.75%)
TYPES	64.84% (-8.32%)	63.86% (-7.89%)
CATEGORIES	62.58% (-10.58%)	71.75% (0.00%)
TEMP	74.72% (+1.56%)	70.43% (-1.33%)
LOC	81.13% (+7.97%)	78.22% (+6.46%)
No SemAbs.	73.16% (0.00%)	71.75% (0.00%)
Majority Class	49.58% (-23.59%)	53.29% (-18.46%)

Though the results are promising, we were interested to get a better understanding of why the trained models work well; thus, we analyzed the rule sets in more detail. In Listing 7.5, an example rule for using all features is shown. The rule shows that

location mentions in combination with incident-related keywords such as "crash" seem to be useful as 139 true positives (TP) and no false positives (FP) are covered. The rule has coverage of 109 TP and 3 FP in 2-C-SEATTLE. Thus, it seems to be a very general rule that is universally applicable.

Listing 7.5: High-quality rule found on tweets of 2-C-MEMPHIS

```
ProperLOC_TFIDF >= 0.029058, TF-IDF <= 1.433658, crashTFIDF  
>= 0.054087, clearTFIDF <= 0.139818 THEN Incident
```

The rule shown in Listing 7.6 is another example of a very general rule (40 TP, no FP in 2-C-MEMPHIS, 294 TP in 2-C-SEATTLE, 39 FP in 2-C-SEATTLE). The rule contains location mentions, incident-related keywords, as well as a LOD feature.

Listing 7.6: Another good rule found on tweets of 2-C-MEMPHIS

```
ProperLOC_TFIDF >= 0.017093, TF-IDF <= 1.75729, carTFIDF <=  
0, trafficTFIDF <= 0.06193, urlTFIDF <= 0.032504, ../  
ontology/AdministrativeRegion <= 0, policeTFIDF <=  
0.080475, DDDTFIDF <= 0 THEN Incident
```

An analysis of the complete rule set shows that LOD features (5 TEMPs), TEMP features (1), and LOC features (5) are part of the rules. Furthermore, the rule covers 20% incident-related instances in the test set compared with not using these features. All features resulting from our semantic abstraction are part of both sets; however, not surprisingly, n-grams are part of the rules that are not present in the other set (12 of 14). Also, the true positive rate is rather high with 85% on the test set.

A manual analysis of a rule part of the model trained on 2-C-MEMPHIS only using LOD features gave us a likely reason for the bad performance of the classifier in 2-C-SEATTLE. The rule contains the LOD features "../yago/YagoPermanentlyLocatedEntity" as well as "../yago/YagoLegalActorGeo" which have to be part of the instance more than once. For 2-C-MEMPHIS, this rule leads to 53 TP (no FP), whereas this rule applied on 2-C-SEATTLE results in 5 TP and a total of 36 FP. Though the rules also contain several TF-IDF features and word-n-grams, a closer look at the LOD features shows that both entities that are indeed representative for incident-related tweets in 2-C-MEMPHIS are indicators for not incident-related tweets in 2-C-SEATTLE. This is an indication that LOD features cannot easily be used and need further filtering before applying a model trained on one city on another one. A further analysis of the rule sets for just using LOD features shows that the coverage drops to 27% (-15.38%), which is an indicator that LOD features useful for 2-C-MEMPHIS are not useful for 2-C-SEATTLE in the feature combinations as they are present in our rules.

The analysis of the rule set for training on 2-C-SEATTLE and testing on 2-C-MEMPHIS shows similar results. For the ALL feature combination, LOD (5), TEMP (1), and LOC (3) features are used in the rule, and all are present in both data sets. In this case, all n-grams are present in the other data set (10 of 10) that are part of the rule. Applying semantic abstraction results in an increase of coverage of the rule set by 14% (61.75% compared with 42.38%), also increasing the true positive rate to 95% (compared with 85%).

The rule shown in Listing 7.7 is an example of a general rule of the ALL feature combination. The rule is applicable for 44 incident-related instances in the training set and applies for 91 instances in the test set without any false positives. Compared with the rule shown in Listing 7.7, similar features seem to be valuable such as TF-IDF scores and the "crash" keyword.

Listing 7.7: A high-quality rule found on tweets of 2-C-SEATTLE

```
TF-IDF <= 1.811512, crashTFIDF >= 0.057693, TF-IDF <=
1.409797, laneTFIDF <= 0.072247 THEN Incident
```

Also in this case, just using LOD features results in a significant drop of coverage to 16.34% (-16.07%) on the test set. The rules indeed show that just one type feature is used. Also, the rule set for using only categories shows that they are not part of the rules trained on 2-C-SEATTLE.

Optimizing LOD Features

As LOD features are valuable for the single-city case, but not directly for the two-city case, we manually tried to conduct a feature selection on these features. For this, we decided to use the most representative LOD features for both data sets. This resulted in eight LOD features that are highly representative for incident-related tweets in both data sets. We confirmed our selection by merging both data sets and calculating the information gain [251] of every single feature, leading to "../ontology/Road", "../ontology/RouteOfTransportation", "../ontology/ArchitecturalStructure", and "../ontology/Infrastructure" as the LOD features part of the top 20 features contributing the highest information gain for the combined data set. They are also part of the eight manually selected features.

Table 49.: F-measures for training on one city and testing on a different city after manual feature selection of LOD features.

	2-C-MEMPHIS to 2-C-SEATTLE	2-C-SEATTLE to 2-C-MEMPHIS
ALL	81.40% (+8.24%)	79.07% (+7.32%)
ALL filtered	73.08% (-0.08%)	76.88% (5.13%)
LOD	64.84% (-8.32%)	64.00% (-7.75%)
LOD filtered	66.57% (-6.60%)	63.86% (-7.89%)
No SemAbs.	73.16% (0.00%)	71.75% (0.00%)

Based on this procedure, we reevaluated the models using only these LOD features. The results presented in Table 49 show that the manual feature selection unfortunately is not valuable. This clearly indicates that more comprehensive methods for feature selection of LOD features are inevitable.

7.5.5.3 Study 3: Summary

In this section, we dealt with the problem of generalizing a classification model in the domain of social media text classification. We showed that semantic abstraction is especially valuable when it comes to training a general model that is trained on one city and applied on a different one. However, using the RIPPER classifier, we could also verify this result when training and testing are done on data from one city. We could further prove the significance of our results with respect to the error rates when semantic abstraction is used.

The results shown above indicate that semantic abstraction is indeed valuable for such types of classification problems. However, a combination of different feature groups seems to be necessary. Just using LOD features tends not to be valuable due to the differences of their occurrences related to incident tweets in the two data sets. Furthermore, the analysis using a rule-based model showed that LOD features are not directly usable for solving the generalization problem as some are representative for incident-related tweets in one data set, but the same features are not representative for the other one. We concluded that LOD features cannot easily be used and need further filtering before applying a model trained on one city on another one. We also conducted a simple manual feature selection but could not improve classification performance.

In comparison with the related work, we showed that using a single approach for creating abstract features, which is mostly done in related work, is not always sufficient. Furthermore, we showed that the combination of multiple approaches is beneficial. Also, we are the first who evaluated an approach on user-generated content that stems from five cities, showing that semantic abstraction is indeed valuable to overcome city-specific features.

7.6 Conclusion

In this chapter, we dealt with the question of how to automatically classify user-generated content related to (small-scale) incidents. To answer this question, we tackled two subordinated questions. As a first step, we conducted a comprehensive feature engineering to find a precise classifier for classifying the type of an incident mentioned in user-generated content. With this classifier, we are able to infer the thematic dimension $R[1]$ automatically on a large amount of data. As a second step, we introduced the novel concept of semantic abstraction, which allows creating general features that help to deal with the special properties of social media data. As a result of the presented approaches, we are able to classify the thematic dimension of a tweet to use this information in the subsequent steps of the framework.

In this chapter, we made the following contributions:

- We introduced an approach for automatically classifying the thematic dimension of user-generated content related to (small-scale) incidents. For this, we conducted a comprehensive feature engineering to find a precise classifier for classifying the type of an incident mentioned in user-generated content. As a result of the engineering steps, we proposed a set of features that are most suitable for classifying the type of an incident. Compared with previous approaches, we are able to precisely classify up to four different classes of incident types.
- We introduced the novel concept of semantic abstraction, which allows creating features that are not city-specific and support training a generalized model. This generalization is important as social media data is heterogeneous and a dynamic source of information. Semantic abstraction makes use of Linked Open Data as an external knowledge base as well as two internal approaches for detecting and abstracting location mentions and temporal expressions. In contrast to previous works, our approach combines several approaches for semantic abstraction and was evaluated on data sets that stem from different cities.
- In an evaluation, we showed that our approach is able to classify tweets with an F-measure of up to 92.10% into four classes. Furthermore, our evaluation showed that the highest F-measure can be achieved using an SVM classifier. The most valuable features we found are word-n-grams in combination with slang and URL replacement, the syntactic feature group, TF-IDF scores, as well as features generated using semantic abstraction.
- Our evaluation results show that semantic abstraction provides much better results with smaller sets compared with not using it. This is an important finding as labeling of data is expensive, and allowing to reduce the amount of data for training helps to save costs.

-
- An in-depth evaluation of semantic abstraction showed that it is especially valuable with respect to creating a general classification model. In emergency management, one would like to reuse one model that is applicable to different data sets (e.g., data from different cities). With our approach, we showed that semantic abstraction provides much better classification results for data sets that stem from a different city. In this case, we are the first who evaluated an approach on user-generated content that stems from five cities, showing that semantic abstraction is indeed valuable to overcome city-specific features.

For future work, data should be collected for more cities to get a better understanding of how our approach behaves for different data sets. Furthermore, additional approaches for semantic abstraction, such as the concept-level abstraction used by [191], could be added. Also, the analysis of the LOD features needs to be extended. For instance, the relation of location mentions and incident-related tweets could be shown and was also visible in the form of LOD features; however, currently, we lack appropriate instruments to make use of this information.

During our research, we found that assigning only one label can result in the loss of important situational information for decision making. Thus, applying multi-label classification on user-generated content might help to address multiple incident types at once. With multi-label learning, a concurrent assignment of all labels can be achieved, which allows a better understanding of the situation at hand. In preliminary works (see [203]), we showed that this is indeed a valuable approach.

As we now know the thematic, temporal, and spatial dimensions of a tweet, we are able to detect incidents. Furthermore, information about the same incident can be clustered. This is shown in the next section.



8 Machine-Based Aggregation of User-Generated Content

In the last chapters, we showed how to derive the thematic [R1], spatial [R2], and temporal information [R3] for each individual information item. In this chapter, we present an approach that uses the inferred information to detect incidents. Furthermore, the approach aggregates all information related to the same incident (see Figure 37). As a result of this step, incidents are detected, and clusters containing information about the same incident are created. Information in these clusters can then be consumed by decision makers to improve their situational awareness.

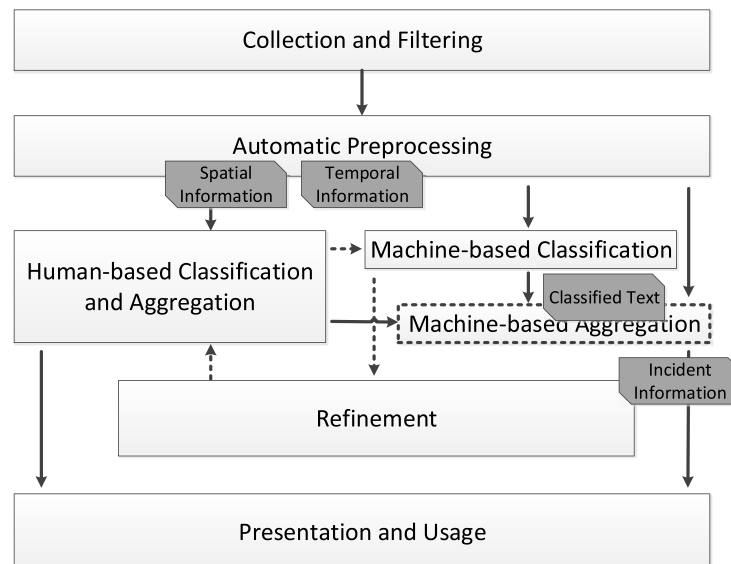


Figure 37.: Machine-based aggregation as a step in the framework.

In this section, we show how we use the individual information items to detect incidents. Furthermore, information items related to incidents need to be identified and clustered. The problem of clustering information related to incidents is shown in Figure 38. To build clusters of reports referring to the same event, it must be decided whether incident reports are about the same incident or about different incidents. For instance, two incident reports might be sent at the same location and might be about the same incident type but were sent at a different point in time. Thus, no aggregation should take place.

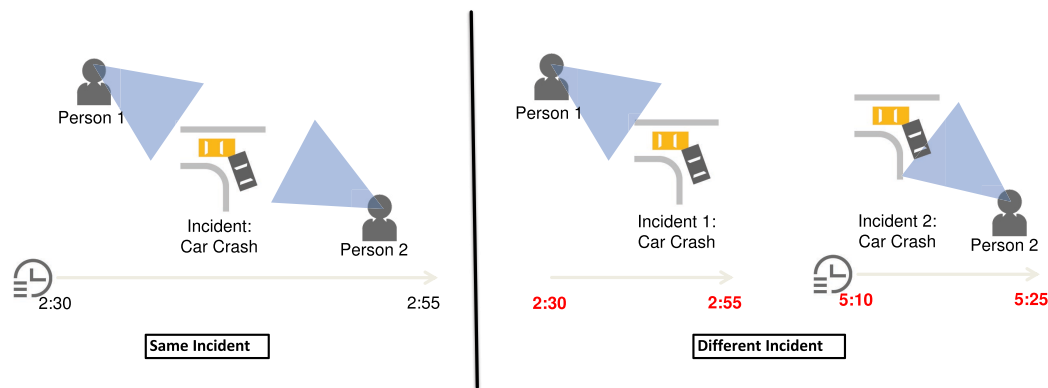


Figure 38.: Two examples showing the clustering of two incident reports. In the first example, the spatial, temporal, and thematic dimensions match; thus, we assume that both incident reports are about the same incident. In the second example, a difference in the temporal dimension is present; thus, both incident reports are most likely about different incidents.

In this section, we need to deal with the following research questions:

- *How can incidents be detected based on large amount of user-generated content?*
- *How can incident reports about the same incident be automatically aggregated?*

A second problem that needs to be addressed is that different emergency management organizations use different vocabularies for incident types. In discussions with police and fire departments in the scope of the research project InfoStrom, we found that the police use a vocabulary of incident types that is defined by the state. For the fire brigades, the situation is different. Here, only guidelines for incident types exist. The guidelines are further detailed by each fire department, according to the local requirements. For example, the police only use the incident type 'Fire', whereas a fire brigade has a much more fine-grained vocabulary to distinguish different types of fire, such as "Roof Truss Fire" and "Forest Fire".

It has been shown that communications issues can derive from these different vocabularies [135], which are critical during rescue operations. Throughout this dissertation, we presented several approaches that might be bound to organization-specific vocabularies of incident types. Thus, incoming incident reports that have been assigned organization-specific incident types need to be aggregated across the different vocabularies. This results in the following question:

- *How can incident reports be automatically aggregated across organization-specific incident type vocabularies?*

The contributions presented in this chapter are the following:

-
- We propose a spatio-temporal-thematic clustering approach, which is able to detect incidents in a large amount of social media data. Furthermore, the approach clusters all information related to the same incident and is able to deal with different incident type vocabularies.
 - We evaluate the approach and show that we are able to detect more than 50% of real-world incidents published in an emergency management system within a whole city. Furthermore, 32.14% of the detected incidents are within a 500 m radius and within a 10 min time interval of the real-world incident, allowing precise spatial and temporal localization. Also, more than 77% of the event clusters created with our approach are indeed related to incidents.
 - We analyze situational information shared in tweets posted in two North American cities. We show that a variety of individuals share information about small-scale incidents. Furthermore, we show that important situational information about affected objects, injured persons, and the location of an incident is shared.

The clustering framework shown in this section is presented as a general framework for different types of events. However, we evaluated the framework for incidents and using social media data.

We present related approaches in Section 8.1, followed by our spatio-temporal-thematic clustering approach for incident detection in Section 8.2. Next, the prototypical evaluation is shown in Section 8.3. Finally, we present an evaluation of our approach with respect to real-world incident data (see Section 8.4).

Parts of this chapter appeared in [199, 195, 201, 197].

8.1 Related Work

Event detection has been addressed widely in the Topic Detection and Tracking (TDT) research [14] and was mainly conducted on structured and long text such as news media [78], [252]. However, in the last years, research also dealt with event detection on short and unstructured texts such as tweets. Thus, the special properties of this type of information needed to be addressed. In the following section, we provide an overview of research focusing on event detection based on unstructured texts.

First, story detection is partly related to our research as new and evolving topics need to be detected [14]. Some approaches focus on summarizing tweets to topics. For instance, [148] apply query expansion to retrieve summaries for events. With this approach, they deal with the problem of evolving topics in tweets by adapting calculating the conditional probability of a term occurring in a query. [156] extended this approach with including the temporal information for generating queries. [12]

further extended the approach with location information. Also, some approaches focus on detecting trends or emerging topics by observing the frequency of how words are used and then grouping these "bursty" words into topics [39], [149]. Also, approaches try to detect related tweets to already-existing events [176], [140]. However, in the following section, we focus only on approaches that deal with detecting new events.

The works presented in this section are differentiated along three aspects: the type of event, the clustering approach, and the metadata used:

- **Type of Event:** Some approaches are not restricted to detect a specific type of event (e.g., when no prior information about an event is known) while others are specialized toward detecting predefined types of events.
- **Clustering Approach:** Different approaches for clustering tweets are employed. While some works focus on unsupervised clustering techniques, other hybrid approaches use supervised learning for classifying clusters or tweets in conjunction with unsupervised clustering.
- **Metadata Used:** The type of metadata used for creating clusters is different for each approach. As we are interested in inferring the spatial and temporal dimensions of an event, we analyze related work with respect to these two dimensions.

In Table 50 and Table 51, overviews of related approaches are given.

Table 50.: Overview of related approaches with respect to the type of event, clustering approaches, and metadata used.

	Type of Event		Clustering Approach		Metadata Used		
	Incident	Other Specific	Unspecific	Supervised	Unsupervised	Temporal	Spatial
Sankaranarayanan et al. [193]			x	x	x	(x)	(x)
Phuvipadawat and Murata [173]			x		x	(x)	
Becker et al. [22]			x	x	x	(x)	
Long et al. [141]			x		x		
Weng and Lee [243]			x		x	(x)	
Li et al. [136]			x		x	(x)	
Parikh and Karlapalem [170]			x		x	(x)	
Cordeiro [54]			x		x		
Gu et al. [90]			x		x	(x)	
Ritter et al. [183]			x	(x)	x	x	
Ishikawa et al. [108]			x		x	(x)	x
Watanabe et al. [241]			x		x		x
Pozdnoukhov and Kaiser [177]			x		x	(x)	(x)
Boettcher and Lee [31]			x	x	x	(x)	x
Xie et al. [247]			x	x	x	(x)	x
Lee and Sumiya [131]		x			x	(x)	x
Chae et al. [41]	x		x		x	(x)	(x)
Agarwal et al. [5]	x			x	x	x	x
Hua et al. [103]	x			x	x		x
Marcus et al. [146]	x	x				(x)	(x)
Jadhav et al. [109]	x				x	(x)	(x)
Walther and Kaiser [238]		x		x	x	(x)	(x)
Li et al. [137]	x			x		x	(x)
Sakaki and Okazaki [192]	x			x		x	x
Our approach	x			x	x	x	x

Table 51.: Overview of related approaches with respect to the approaches used for event detection and clustering.

	Supervised Approach	Unsupervised Approach	Event Scale
Sankaranarayanan et al. [193]	NB	K-Means	large
Phuvipadawat and Murata [173]		Single-Pass Clustering	large
Becker et al. [22]	SVM	Single-Pass Clustering	large
Long et al. [141]		Hierarchical Clustering	large
Weng and Lee [243]		Wavelet Analysis and Graph Partitioning	large
Li et al. [136]		Hierarchical Clustering	large
Parikh and Karlapalem [170]		Hierarchical Clustering	large
Cordeiro [54]		Wavelet Analysis and LDA	large
Gu et al. [90]		Hierarchical Clustering	large
Ritter et al. [183]	CRF	LDA Topic Modeling	large, (small)
Ishikawa et al. [108]		Single-Pass Clustering	large, (small)
Watanabe et al. [241]		Spatial Clustering	small
Pozdnoukhov and Kaiser [177]		LDA Topic Modeling, Statistical Burst Detection	large, (small)
Boettcher and Lee [31]	Logistic Regression	Density-based Clustering	small
Xie et al. [247]	SVM, NB, Log. Reg.	Statistical Modeling	small
Lee and Sumiya [131]		Statistical Modeling	small
Chae et al. [41]		LDA Topic Modeling	large, (small)
Agarwal et al. [4]	NB + SVM	Locality Sensitive Hashing	small
Hua et al. [103]	not defined	Graph Partitioning	small
Marcus et al. [146]		Statistical Burst Detection	small
Jadhav et al. [109]		Spatio-Temporal Clustering	large, (small)
Walther and Kaiser [238]	Decision Trees (J48)	Spatio-Temporal Clustering	small
Li et al. [137]	Not Defined	Not Defined	small
Sakaki and Okazaki [192]	SVM	No Clustering	large
Our approach	SVM	Spatio-Temporal-Thematic Clustering	small

8.1.1 Type of Event

Regarding the type of event, the approaches focus on either unspecified events or on predefined event types. We show that several approaches focus on detecting events without an explicit event type, such as trends or breaking news. In contrast, only few approaches focus on detecting events of a specific type. Also, small-scale incident detection is rather rare as a use case.

Unspecific Events

Sankaranarayanan et al. [193] presented TwitterStand for late breaking news detection. For prefiltering all incoming information, tweets are classified into two classes ("news" and "junk") using an NB classifier. Based on the preclassified information, online clustering is applied for topic detection. For this, TF-IDF scores are used for calculating the cosine similarity between the tweets. Furthermore, their approach makes use of temporal and spatial information. Temporal information is used for outdated clusters (e.g., after three days) so that new clusters can be created. Spatial information is extracted for each cluster by using geolocation information from each tweet. Unfortunately, no evaluation was provided by the authors.

The approach presented by Phuvipadawat and Murata [173] first applies keyword-based filtering for retrieving relevant samples of tweets. For grouping messages, TF-IDF scores are calculated and weighted if a term is a proper noun. Based on these, the similarity of messages is calculated, and online clustering is applied. Temporal information is used for ranking clusters according to their newness. Also, no evaluation was provided.

Becker, Naaman, and Gravano [23] presented a system for event detection. Based on the cosine similarity of the TF-IDF scores of each tweet to a cluster, an online clustering is performed [22]. Afterward, each cluster is assigned a label whether it is an event cluster or not. For this, several Twitter-specific features (see Section 7.2) are used, which includes temporal information. An evaluation was conducted on 374 manually annotated event clusters consisting of tweets from New York City and shows an F1 score of 83.7% for testing.

The approach of Long et al. [141] first determines word cooccurrences that are representative for an event. Based on these, a cooccurrence graph is built. Hierarchical divisive clustering is applied on the graph to determine event clusters. As a last step, the documents belonging to one cluster are summarized using cosine similarity between messages to retrieve the most relevant posts describing the clusters. In an evaluation, the authors showed that hierarchical divisive clustering outperforms k-means and traditional hierarchical clustering. Furthermore, they showed that their approach is able to achieve a precision of 40% if 20 event clusters are created.

Weng et al. [243] used "signals" for individual words to capture the bursts in the appearance of words. For this, wavelet analysis is used on time-dependent TF-IDF scores (DF-IDF). Each word is then represented as its corresponding signal. Finally, events are clustered based on words with similar burst patterns. An evaluation was conducted on 21 event clusters, showing a precision of 76.2%. A similar approach is followed by Cordeiro [54], who also employs wavelet analysis, but this time on hashtag occurrences. Thus, an increase on the mention of a hashtag is used for event detection. If a peak in the use of hashtags is detected, LDA is applied to extract a set of topics to provide a summary of the event description.

Li, Sun, and Datta [136] proposed an approach that first detects bursty tweet segments, which are nonoverlapping phrases of word-n-grams. From these, bursty segments with increased frequency of tweets are detected in a time interval. Next, the similarity between the segments is calculated using cosine similarity on TF-IDF scores to cluster segments to events using k-nearest neighbor algorithm. The approach was compared against [243] and showed better performance (precision of 86.1% and a recall of 75 distinct events).

The approach presented by Parikh and Karlapalem [170] also relies on detecting bursty keywords in a time interval. For detecting event clusters, hierarchical clustering is used on the keywords. Compared with previous works of [243] and [136], they try to reduce the computational complexity by optimizing the temporal segregation. The authors reported a precision of 91% for detecting 23 events.

Gu et al. [90] proposed ETree for summarizing information about events. First, messages for an event are collected using keyword-based search. Second, frequent word-n-grams are detected and used for grouping event-related messages. A similarity is calculated using the cosine similarity on TF-IDF scores and is clustered accordingly. Furthermore, temporal information is used to detect the relationships between clusters.

Ishikawa et al. [108] tried to identify trending topics in geographically restricted areas. For this, geotagged tweets excluding tweets from location-based services are used. Next, word relations are identified using Wikipedia concepts. Incremental clustering is applied using text similarity (unclear how calculated). Third, bursty topics are detected by comparing the frequency of tweets and the overall number of tweets in a cluster. Also, no evaluation was provided.

Pozdnoukhov and Kaiser [177] applied LDA topic modeling on geotagged tweets. The resulting clusters are then clustered using kernel density estimates, which is a nonparametric statistical approach [215]. The number of tweets per hour is then observed for irregularities to identify events. Except for a case study, no formal evaluation was conducted.

Ritter et al. [184] presented TwiCal for event detection and classification. First, named entities and temporal expressions are extracted from tweets. Second, part-

of-speech tagging is applied. Based on the identified named entities and POS tags, event phrases are detected using a Conditional Random Fields approach for tagging (F-measure of 0.64%). Finally, the extracted events are clustered in an unsupervised manner using LDA and types are inferred using Gibbs sampling [89]. An evaluation showed that their approach is able to correctly classify (precision of 0.7%) for 100 events and (precision of 0.34%) for 1,000 events.

The approaches presented so far were applied for large-scale event detection; however, some also tried to focus on small-scale event detection by taking spatial information into account. For instance, Lee and Sumiya [131] try to detect the occurrence of local events such as local festivals. In contrast to other works, regions of interest are created using geotagged tweets and applying k-means clustering on the geographical regions the tweets were sent from. Then behavior patterns are mined for each user by monitoring tweets in these regions. The patterns are calculated for six-hour time intervals and created using the number of tweets, the number of users, and the movement between geographical regions. Unusual behaviors are then detected using statistical models. If irregularities are detected, they are regarded as events. Though the approach showed good performance for detecting new events, precision was very low when detecting 15 events (1.8%).

Boettcher and Lee [31] presented another approach for local event detection. First, tweets are crawled and preprocessed. Next, density-based clustering [69] is applied for fixed time intervals to identify event clusters. This approach clusters tweets according to their spatial and thematic proximity (i.e., common keywords). For spatial proximity, only geotagged tweets are used. Finally, the resulting clusters are labeled into "local event" or "no local event" using a logistic regression classifier. An evaluation showed that the approach is capable of detecting local events with 68% precision on 74 local events.

Xie et al. [247] proposed to combine time series prediction and a supervised classifier for small-scale event detection. First, the statistical distribution data for a certain region and time interval is observed using Gaussian Process Regression (GPR) [244]. Outliers are regarded as potential events, which are then labeled. The classifier was built using textual features as well as features derived of the GPR approach. In an evaluation, the authors showed that they are able to classify event clusters with a high accuracy of 89%. Though they show examples for detected real-world events, they do not report any evaluation results regarding this aspect.

Jasmine, a system for detecting small-scale events, was presented by Watanabe et al. [241]. For this, spatial and thematic information is used. Tweets are constantly crawled and geotagged using a lightweight geolocalization approach. Based on the frequency of tweets in a certain area, popular regions are detected; however, it is not described how popular regions are differentiated from regular regions. For these

regions, key terms, which are terms that appear three times or more, are extracted. An evaluation showed that 25.5% of 346 detected events are local events.

Specific Events

In contrast to the approaches designed for detecting unspecific events, other approaches focus on detecting specific events such as incidents. For this, location and temporal information is used for restricting the area for event detection. Furthermore, as we show, supervised learning is applied to classify the type of event.

Chae et al. [41] presented a spatio-temporal approach for detecting events. They use a Latent Dirichlet Allocation (LDA) topic modeling approach [29], which is an unsupervised approach to identify latent topics and corresponding clusters. To detect irregular events, seasonal trends are tracked, and abnormal events are identified based on the frequency of messages per topic. In contrast to other works, a user can specify a temporal and spatial boundary for filtering events. Furthermore, the number of resulting events needs to be specified too. The approach was applied on large-scale incidents such as earthquakes and showed that important clusters can be found. However, no formal evaluation was provided.

Li et al. [137] presented TEDAS, which uses spatial and temporal information for detecting incident-related events. First, crawled tweets are classified using a supervised learning approach. Next, temporal and spatial information are extracted from the tweet's metadata. For geolocating a tweet, GPS information from the user's social network is used. Based on these information, tweets are clustered and ranked; however, it remains unclear which approach is used for clustering. Also, the system was not evaluated.

Agarwal et al. [5] proposed to first use a supervised approach for classifying tweets according to event types. Next, clusters are built based on Locality Sensitive Hashing (LSH) [181]. Temporal and spatial metadata is then used to provide information about the clusters. Though they were able to detect location mentions in 74% of the cases, an evaluation of the accuracy is not provided. Furthermore, no information about the temporal extraction is given. They reported that they are able to detect events with an accuracy of 76.6%, although it is not described what is used as ground truth for events.

Hua et al. [103] presented STED, a system for small-scale event detection. Graph partitioning [161] is applied for identifying and clustering similar tweets. Then the resulting clusters are labeled using a not-further-specified supervised approach. The approach was tested on (an unknown number of) tweets collected in Latin America and shows a precision of 72% and a recall of 74% for event detection.

TwitInfo [146] is a system for summarizing tweets related to user-defined keywords for an event. The authors introduced a peak detection algorithm that continually

tracks the whole Twitter stream and identifies outliers depending on the frequency of the keywords detected in the tweet messages, which they regard as events. The system was evaluated on sport events and earthquakes, reporting a precision of 77% and a recall of 77% for detecting sports events and a precision of 14% and a recall of 100% for detecting major earthquakes.

Jadhav et al. [109] presented an approach using Twitter for sense making (e.g., for the identification of events) in tweets. Incoming tweets are analyzed based on spatial, temporal, and thematic dimensions. These dimensions are used to cluster tweets according to events. For this, the user can specify a spatial and temporal bounding box for aggregating tweets. Based on this, descriptions of the events are generated using TF-IDF scores of word-n-grams. The five n-grams with the highest score are chosen as descriptions. Though the platform follows a similar approach as this dissertation, no evaluation results are available. Furthermore, the authors admit that the spatio-temporal-thematic analysis is done with a week of lag.

Walther and Kaisser [238] presented a similar approach for small-scale event detection. Tweets are constantly crawled and clustered according to their spatial and temporal proximity as well as their frequency. For example, at least three tweets that were created in the last 30 min in a 200 m radius form a new cluster. Their approach relies only on geotagged tweets, and no information about what is used for temporal boxing is provided. The resulting clusters are finally labeled into two clusters ("real-world event" or "no event"). No evaluation was conducted with respect to their capabilities of detecting real-world events.

Sakaki et al. [192] used an SVM classifier to detect earthquakes as a type of large-scale incident. The creation date and inferred spatial information is then used for estimating the earthquake center. However, their approach is restricted for detecting only one event and not multiple events simultaneously. In an evaluation, they showed that their system is able to detect 10 of 10 earthquake events.

8.1.2 Clustering Approach

For detecting incidents and aggregating related information to clusters, different approaches are employed. While some works focus on unsupervised clustering techniques, other hybrid approaches use supervised learning for classifying clusters or tweets in conjunction with unsupervised clustering.

Unsupervised Approaches

Various unsupervised clustering techniques are used for handling large volumes of social media data. Clustering implementations of partitioning algorithms such as k-means and hierarchical approaches (see [110] for detailed explanations) are com-

monly used for event detection [193, 141, 136, 170, 90]. One major disadvantage of these approaches is that the number of clusters needs to be specified beforehand. However, for incident detection, we have no a priori knowledge about the number of clusters. Also, when hierarchical clustering is used, a threshold needs to be specified to define the distance between the closest clusters, which needs to be tuned. Furthermore, the traditional implementations are typically slow for a large amount of data. To overcome this limitation, single-pass clustering approaches, such as leader-follower clustering, were applied [173, 22, 108]. These approaches have the advantage that each instance is only processed once for calculating the similarity to all clusters. Though this approach is simple to use, it is dependent on the order the data [75]. Furthermore, it tends to produce large clusters.

Graph partitioning [120] is another branch of clustering algorithms applied for event detection [103, 243]. However, as this approach tries to balance partitions, it is not suitable for clustering events in social media data [21].

[31] applied density-based clustering using DBScan [69], which is a clustering according to the density of instances (e.g., based on their spatial proximity). Nevertheless, the minimum number of instances contained in a cluster as well as a threshold needs to be specified as parameters [117], which are not known in advance.

Agarwal et al. [5] applied LSH [181], which is implemented as the k-nearest neighbor search on all instances using hashing functions. Also in this case, the parameters for the hash functions and the similarity functions need to be optimized to achieve good results.

In contrast to these clustering techniques, topic modeling was used for aggregating related information. These approaches directly use the tweet's text to create topic models (e.g., based on LDA) [184, 177, 41, 54]. Also, the text is used to create statistical models [131, 146] (e.g., for burst detection). However, for small-scale incidents, these approaches are not applicable as the number of tweets shared for small-scale incidents is rather low. Furthermore, spatial and temporal information remains unused.

Compared with these traditional approaches, some approaches were developed to use metadata during the clustering process. [241] made use of the geotags and applies simple spatial clustering. In contrast to other works, [109] and [238] followed a spatio-temporal clustering approach. For this, a spatial and temporal bounding box was defined to which extent newly incoming information is aggregated. However, both approaches used thematic information after clustering.

Hybrid Approaches

A common problem of unsupervised approaches is that the clusters mostly remain unlabeled; thus, no thematic information describing the type of event is assigned.

For labeling clusters, the combination of unsupervised clustering with supervised classification approaches was conducted in rare cases. This is due to the fact that supervised approaches are only necessary if specific event types such as incidents need to be detected or which types are known beforehand. In this case, techniques such as NB, SVMs, or decision trees are used for classification. Some approaches classify tweets before aggregating information, while others focus on classifying the event clusters. We prefer the first approach as it reduces the computational overhead as the number of tweets can be significantly reduced; thus, clustering and preprocessing steps are much faster.

A disadvantage of applying supervised approaches for event classification is that Twitter is an evolving environment, which means that different terms are used and these terms are location dependent. Thus, these approaches suffer the fact that a classifier that is not generalizable is used. We already dealt with this problem in Chapter 7 and showed that we are able to handle data from different cities. Furthermore, in Chapter 9, we show that refining a classifier can be achieved with limited effort. This is why we follow a hybrid approach, combining supervised classification and unsupervised clustering.

8.1.3 Metadata Used

The use of metadata for clustering varies considerably across the related approaches. The majority of the presented related work mainly relies on the thematic dimension of tweets (i.e., tweets are classified based on text similarity). For calculating this similarity, TF-IDF representations of the tweets contained in the clusters or the tweets prior clustering are used in conjunction with Twitter-specific features such as hash-tags. Based on traditional distance metrics such as Euclidean distance, Pearson's correlation coefficient, or cosine similarity, the distance is computed. In contrast, with our approach, we follow the idea that only a single incident type is provided instead of a summary based on all tweets contained in a cluster; thus, allowing the fast understanding of the situational picture. A more sophisticated summary would be necessary for large-scale incidents, whereas for small-scale incidents, the number of tweets is rather low and can easily be interpreted manually.

Temporal metadata is used for detecting changes in the frequency tweets are created or words are used. Some approaches also include temporal metadata in the clustering process ([184, 4, 137, 192]). However, except [184], none of the presented approaches make use of temporal expression recognition.

As we have shown, spatial information is mainly used by approaches that focus on detecting small-scale events. Nevertheless, some approaches only use explicitly geotagged tweets ([177, 41, 146, 109, 238, 137]) while others propose approaches

for geotagging ([108, 241, 31, 131, 4, 103, 192]). The former case allows higher precision, while in the latter case, more information is available.

8.1.4 Summary

In the following, we discuss related approaches.

Type of Event: As shown in this section, related approaches focus on either unspecified events or on predefined event types. Furthermore, these approaches can be differentiated with respect to the scale of event. Approaches coping with unspecific event detection are mostly focused on large-scale events, whereas specific event types are detected on small scale. In particular, small-scale incident detection is a rare use case in the domain of event detection.

Clustering Approaches: As shown in this section, clustering-based approaches are frequently used for event detection. However, there might not be an obvious clustering of unknown data; thus, depending on the approach that is applied, clustering might be performed at various levels of granularity. In most cases, the number of optimal clusters is unknown and has to be determined with high computational costs. Also, thresholds need to be tuned, which is not always feasible when it comes to real-time detection.

Another major issue with related approaches is that unsupervised clustering is applied on all data that is available, and labeling of the thematic dimension is conducted based on the created clusters. Though this approach is more robust, processing and clustering millions of tweets are rather time-consuming.

In contrast to this, we propose a straight forward clustering of unknown event-related data and directly deal with the problem of finding the number of optimal clusters. For this, we follow a spatio-temporal clustering approach as it is also used by [109] and [238]. However, compared with these works, we apply classification before clustering, thus allowing to discard irrelevant content prior to building clusters. As a result of our approach, less noise is contained in the clusters, making them more manageable for decision makers.

Metadata Used: The type of metadata used for creating clusters is different for each approach. The majority of related works rely on the thematic dimension of tweets (i.e., tweets are classified or aggregated based on text similarity). Other metadata is seldom taken into account.

Temporal information is used for detecting changes in the frequency that tweets are created or words are used. Furthermore, spatial information is mainly used by approaches that focus on detecting small-scale events. However, in most related works only explicitly geotagged data is used; thus, important information is not

taken into account and discarded. Also, the approaches focusing on incident-related events do not take organization-specific vocabularies into account.

In contrast, we apply geotagging as well as temporal extraction for identifying more precise incident clusters. Also, different vocabularies can be handled with our approach.

Evaluation Methodologies: A major issue we found is that evaluations are mostly missing or are done on small subsamples for which no public data sets are available. Although there are aims to provide publicly available data sets [153], no comparison of related works with respect to a common data set is possible. Furthermore, the two approaches related to this dissertation do not provide any evaluation results (see [109] and [238]). However, we follow an approach that compares the results to real-world incident information published as Open Data, which is freely available to other researchers.

8.2 Approach

In the following section, we present our approach for detecting incidents using user-generated content based on spatial, temporal, and thematic information. Furthermore, we show how we cluster incident reports about the same incident. In contrast to previous works, we make use of all metadata available, thus allowing high coverage for detecting small-scale incidents. Furthermore, our approach is a hybrid approach making use of supervised learning for filtering out noise and unsupervised clustering.

We assume that every incident report is either related or not related to a specific real-world event. Thus, we propose to cluster all reports based on the three dimensions that we inferred: temporal and spatial dimensions as well as the incident type. As we determined all of these dimensions throughout the processing steps shown in this dissertation, each report can be aggregated to a cluster.

In the following, we first present the general approach for incident detection and the clustering of related information. Next, we introduce how clustering across different vocabularies is performed. As different incident reports with contradicting information are clustered, we finally deal with merging incident reports.

8.2.1 Incident Detection and Clustering of Related Information

Detecting incidents using incident reports is based on the previously described steps of the processing pipeline. In Figure 39, it is shown how the information is finally used for clustering information related to the same incident. In general, the approach comprises three steps and works as follows:

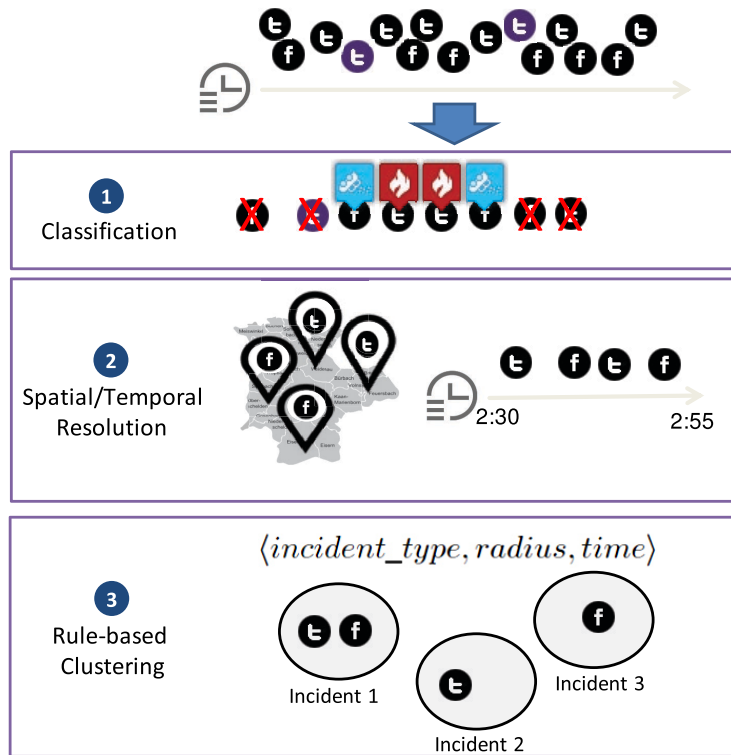


Figure 39.: Approach for detecting incidents and clustering incident reports.

Classification of Incident Types

First, the thematic dimension of an information item is determined (see Figure 39, ①). In our case, an incident type needs to be assigned to the reported text. The thematic dimension is either assigned manually (e.g., to a text using human-based classification; see Chapter 6), or it is inferred for a text using automatic classification techniques (see Chapter 7). As a result of this, incident reports are identified in large amounts of potentially available information items. Assigning the incident type before conducting further processing steps such as spatial or temporal resolution allows filtering out irrelevant information, significantly reducing the amount of data that needs to be processed in the following steps.

Furthermore, when classifying incident types, the organization-specific vocabularies need to be regarded. In discussions with police and fire departments, we were provided with hierarchical ordered incident types. For example, the fire brigade has different subtypes for "Fire" such as "Roof Truss Fire" and "Forest Fire". To deal with these vocabularies, *subtype relations* are assigned to an incident report. The subtypes are determined based on vocabularies predefined by emergency organizations, and the assignment is done within vocabularies. As a result of this, a subgraph comprising all incident types is used as a thematic dimension for an incident report.

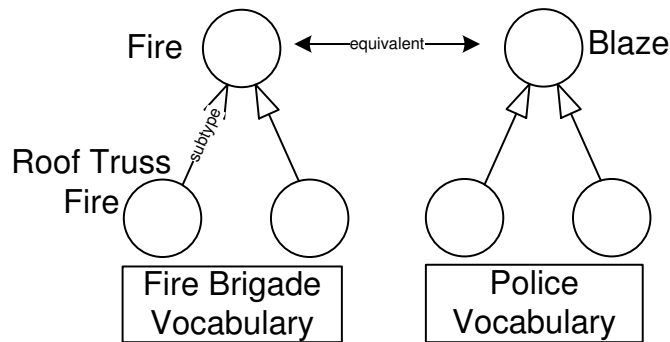


Figure 40.: Assignment of incident types within and across incident type vocabularies.

In addition to this, we take *equivalence relations* across types into account. For instance, the incident type "Blaze" in the police vocabulary is equivalent to the type "Fire" in the fire brigade vocabulary. If these equivalence relations are provided, we also assign the respective subgraph across vocabularies as a thematic dimension for an incident report.

In Figure 40, an example of this is shown. An incident report with the incident type "Fire" is assigned subtypes and equivalent types within and across vocabularies of different organizations. The subtype relation is asserted between the types of "Fire" and "Roof Truss Fire" in the fire brigade vocabulary, indicating that a roof truss fire is a special type of fire. An equivalence relation is used to establish equivalent types across vocabularies. For example, it is indicated that the type "Blaze" in the police vocabulary is equivalent to the type "Fire" in the fire brigade vocabulary. Thus, any incident report of the type "Fire" is automatically declared an incident of type "Blaze" and vice versa. However, it is necessary that these subtypes and equivalence relations are predefined by the respective organizations.

Spatial and Temporal Resolution

As a second step, spatial and temporal information is extracted using the mechanisms described in Chapter 4 and Chapter 5, respectively (see Figure 39, ②). As a result of this, each incident report is assigned a precise location in the form of a GPS coordinate pair. Furthermore, a date that specifies the starting date of the incident described in the incident report is determined.

Rule-based Detection and Clustering

Third, a rule-based approach is used for detecting incidents and creating clusters of related information (see Figure 39, ③). An incident described by a newly incoming

incident report is compared with the previously reported incidents. For the outcome of this comparison, we differentiate two cases for the *matching* of incident reports and previously reported incidents:

- No match to existing incident: if no existing incident could be detected, a new *incident cluster* is created comprising the spatio-temporal and thematic information of the incident report.
- Match to existing incident: if the newly reported incident lies within the spatial, temporal, and thematic extent of an existing incident, then the incident report is aggregated with the existing incident clusters.

To conduct this matching, we follow a rule-based approach. A rule specifies the spatial, temporal, and thematic extent used to assert the equivalence of a new incident report and an existing incident cluster. A new incident report is aggregated with an existing incident if its spatial, temporal, and thematic location falls within the extent that the rule asserts to the existing incident. These rules are described as a triple of the following form:

$$\langle incident_type, radius, time \rangle \quad (21)$$

The spatial extent is a radius in meters drawn around the spatial location of the incident. The temporal extent is a time span in minutes calculated from the creation time of the initial incident. The thematic information is referenced by specific vocabularies that define the possible type of incidents. In our case, the thematic extent is the subgraph indicated by the type given in the rule. All types that are subtypes according to the type assertions belong to the thematic extent of the incident.

The following rule is an example rule for our approach:

$$\langle Car_Crash, 50\,m, 30\,min \rangle \quad (22)$$

The rule asserts that each reported incident of the type "Car Crash" or possible subtypes like "Car Crash with Injured Persons" is identical to a previously reported incident if it is of the same type, within a range of 50 m and within a time of 30 min. Thus, the corresponding tweet is assigned to the incident cluster.

In Listing 11, the algorithm for our approach is shown.

Algorithm 1: Algorithm for rule-based clustering.

Data: Incident reports IR
Result: Incident clusters C

```
1 forall the incident reports  $ir$  in  $IR$  do
2   Get all rules  $R$  applicable for incident types of  $ir$ 
3   forall the rules  $r$  in  $R$  do
4     Get all existing incident clusters  $C$  within time interval of  $r$ ;
5     forall the clusters  $c$  in  $C$  do
6       if  $Distance(c, ir) < radius$  of  $r$  then
7         if incident type is class or subclass of rule types then
8           Apply Rule-based Clustering ;           /* see Sec.8.2.1 */
9           Reevaluate rules for new incident;
10        else
11          Create new incident cluster  $c$ ;
```

Summary

We presented a spatio-temporal-thematic clustering approach that is able to detect incidents in a large amount of social media data. Furthermore, we showed how the approach is applied for clustering incident reports related to the same event. Also, we showed how different vocabularies are taken into account.

As the temporal expression recognition and our approach for geolocalization are prone to errors, the clustering approach needs to deal with incomplete information. In our case, spatial information is replaced with a common spatial center (e.g., the center of the city for which the tweets are used). Missing temporal information is replaced with the creation date of the tweet. Thus, even with one or two missing dimensions, we are still able to build clusters. However, this may lead to a variety of clusters containing incident-related tweets; thus, further manual filtering might be needed. However, important information is not lost as it is assigned at least an incident type.

With the help of these three steps, a rule engine can compute whether incident reports are clustered as they describe the same event. We argue that these rules should be specified by emergency managers to match incident types according to their needs. However, this specification requires domain knowledge in emergency management, although it would be possible to automatically derive optimal rules based on real-world data. Nevertheless, as these rules comprise different vocabular-

ies only known by emergency staff, automatically inferred rules might need manual postprocessing.

8.2.2 Clustering Across Different Vocabularies

As described before, one of our goals is to match incidents across and within vocabularies of emergency management organizations. For this, the types of the newly created incident reports and the existing clusters are compared. The comparison of incident types requires the integration of different vocabularies and the consideration of the subsumption and equivalence relations within and across vocabularies of different organizations. For matching across incident types, we identified four cases illustrated in Figure 41:

1. The newly reported incident is of the same type as an existing incident. This case is trivial, and the clustering takes place.
2. The newly reported incident is of a more specific type than an existing incident. In this case, the incident report is aggregated to the cluster if the rule applies to the more general type of the existing incident cluster.
3. The newly reported incident is of a more general type than an existing incident. The incident report is aggregated to the cluster if the rule applies to the more general type of the newly reported incident; otherwise, no aggregation takes place.
4. The newly reported incident is of a type that is neither a more general type nor a more specific type than the type of an existing incident cluster. This case can appear in two variants. If the types of both share a common more general type to which the match assertion applies, then the incident report is matched to the cluster. Otherwise, they are regarded as two distinct incidents; thus, a new incident cluster is created.

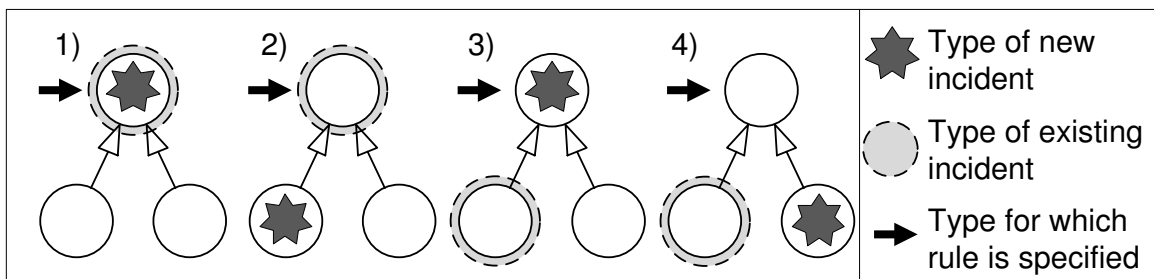


Figure 41.: The four cases when the thematic extent of an existing incident matches a newly reported incident. The numbers correspond to the enumeration in the text above.

Based on these cases, our rule-based approach is able to aggregate incident reports across organization-specific vocabularies. This allows a much better communication across different emergency management organizations.

8.2.3 Aggregation of Incident Reports to Incident Clusters

In the last sections, we described how the clustering of incident reports is conducted. While aggregating reports to clusters, new spatial, temporal, and thematic information needs to be merged to the existing one. In this section, we describe how this merging is performed.

Spatial Merging: In the case of a successful match between an incident report and an incident cluster, the spatial information of the incident report and the incident cluster are merged. The spatial locations of the incident reports are merged to a weighted mean location. The weights take the ratio of $1 : N$ between the new incident report and the incident reports in the cluster where N is the number of reports that are already associated with the existing incident cluster. This ensures that all incident reports have the same impact on the mean incident location.

Temporal Merging: The temporal dimension is not merged, but the time of the first incident report indicates the starting time of the incident a cluster is referring to.

Type Merging: When merging incident types, we aim to maximize the information available for the merged incident. This means that the type after the merge should not be more general than the type of the incident before the merge. For example, if the type "Roof Truss Fire" is reported, and an incident of the more general type "Fire" matches the spatio-temporal extent and a rule matches incidents of type "Fire", then the existing incident of type "Fire" is re-typed as "Roof Truss Fire". If no subtype relation holds between the types of the incidents, then the intersection of the types is taken as the new type. The intersection expresses that the incident has both types. For example, if an incident of the type "Car Crash with Injured Person" spatiotemporally matches an incident of the type "Car Crash with Perpetrator on Site" and these two types are both subtypes of the same type "Car Crash" to which the matching rule applies, then the incident will be of both types, indicating that there are injured persons and that the perpetrator is on-site. This allows keeping the specific information that different reports provide.

8.3 Prototypical Realization

In the following, we provide details about the prototypical implementation of the described approach.

Creating Rules

The creation of new rules for applying the rule-based clustering is done in the rules view. Users can define rules that help identifying whether two incident reports describe the same incident (e.g. *if there is an incident of type "Car Crash" reported within 50 m radius and less than 30 min after another report of "Car Crash", then this must be the same incident*).

Home Incidents Rules About

Create Rule

New Rule

Radius (m):
150

Incident Types:

- ☒ http://infostrom.sap.com/ontology/firefighterincidenttypes.owl#Incident
 - ☒ http://infostrom.sap.com/ontology/firefighterincidenttypes.owl#ABC
 - ☒ http://infostrom.sap.com/ontology/firefighterincidenttypes.owl#Bombe
 - ☒ http://infostrom.sap.com/ontology/firefighterincidenttypes.owl#Einsturz_1
 - ☒ http://infostrom.sap.com/ontology/firefighterincidenttypes.owl#Einsturz_2
 - ☒ http://infostrom.sap.com/ontology/firefighterincidenttypes.owl#Explo
 - ☒ http://infostrom.sap.com/ontology/firefighterincidenttypes.owl#Fahrzeug
 - ☒ http://infostrom.sap.com/ontology/firefighterincidenttypes.owl#LKW_1
 - ☒ http://infostrom.sap.com/ontology/firefighterincidenttypes.owl#LKW_2
 - ☒ http://infostrom.sap.com/ontology/firefighterincidenttypes.owl#PKW_1
 - ☒ http://infostrom.sap.com/ontology/firefighterincidenttypes.owl#PKW_2
- ☐ http://infostrom.sap.com/ontology/firefighterincidenttypes.owl#Unfall_1
- ☐ http://infostrom.sap.com/ontology/firefighterincidenttypes.owl#Unfall_2

Duration in Minutes:
60

Create

[Back to List](#)

Figure 42.: The user interface to create rules. The vocabulary contains the German terms for the fire brigade involved in the InfoStrom project.

In Figure 42, the creation of a new rule based on the fire brigade vocabulary is shown⁵². The vocabulary is presented based on the organization the authenticated user belongs to. For creating rules, the user defines the radius and the temporal extent of the incident. Furthermore, a set of incident types can be chosen to avoid the creation of redundant rules.

⁵² The respective vocabularies have been created in the context of the research project InfoStrom funded by the BMBF (13|10712).

The vocabularies are implemented in RDF Schema⁵³ (RDFS). RDFS specifies how to implement vocabularies in the RDF⁵⁴. In RDF, each incident type is modeled as a class. RDFS allows us to express lightweight semantic statements that restrict some interpretations of classes, for example, by specifying class hierarchies with the *rdfs:subClassOf* property. RDF vocabularies are graphs of classes connected by properties. We use the *rdfs:subClassOf* property to structure the rather long lists of incident types used by the police and the fire brigade.

The cases for matching on the vocabulary level is implemented in a separate RDF file, which makes use of the organizational vocabularies. The cases are exclusively implemented through *rdfs:subClassOf* properties. The equivalence of two incident types T_1 and T_2 is implemented by asserting that T_1 is a subtype of T_2 and vice versa that T_2 is a subtype of T_1 . In RDF, this asserts an extensional equivalence, which means that both types have exactly the same individual incidents. Therefore, any incident of type T_1 is automatically interpreted as an incident of type T_2 as well. For example, the incident type "Fire" in the fire brigade vocabulary is asserted to be equivalent to the incident type "Blaze" in the police vocabulary. When a policeman reports an incident of type "Blaze", the fireman would be able to find it as a "Fire" incident in his terms.

SID: Small-Scale Incident Detector

For realizing the framework in a prototypical implementation, we developed the Small-Scale Incident Detector (SID). The application constantly collects information from social media platforms such as Twitter and stores all information in a database. Each information item is then processed and classified according to the incident types. The resulting incident reports are then spatio-temporally localized so that the matching can be applied. The application constantly applies the rule-based clustering if a new incident report is retrieved.

As shown in Figure 43, the application provides an overview of all detected incidents within a city. For this, information regarding the three dimensions that build the fundamentals of this dissertation is shown: the incident type, the incident date, and the location. Furthermore, based on the individual classification confidences of each incident report, an aggregated confidence score is calculated for each incident cluster to indicate how probable each prediction is. This way, decision makers are provided with an aggregated and easy-to-consume overview of all incidents that could be detected in user-generated content.

⁵³ See <http://www.w3.org/TR/rdf-schema/> [Accessed: 01.02.2013]

⁵⁴ See <http://www.w3.org/RDF/> [Accessed: 01.02.2013]

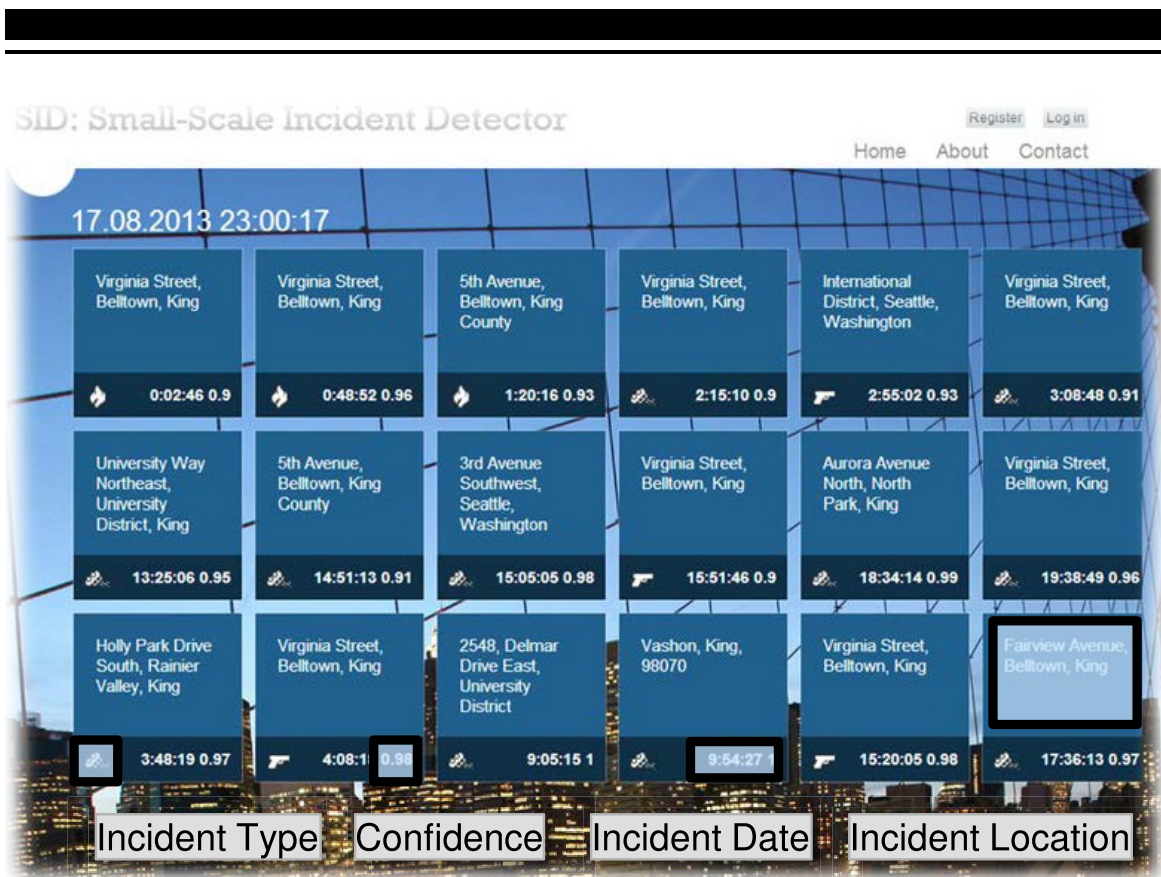


Figure 43.: SID as an application for incident detection based on user-generated content.

In Figure 44, an example of information is displayed for each incident cluster is shown. The application provides an overview of all tweets that were aggregated. Information about the incident type and the individual confidence score for predicting the incident type is shown. Using SID, decision makers have a single access point that allows easy use of incident information derived from social media data for decision making.

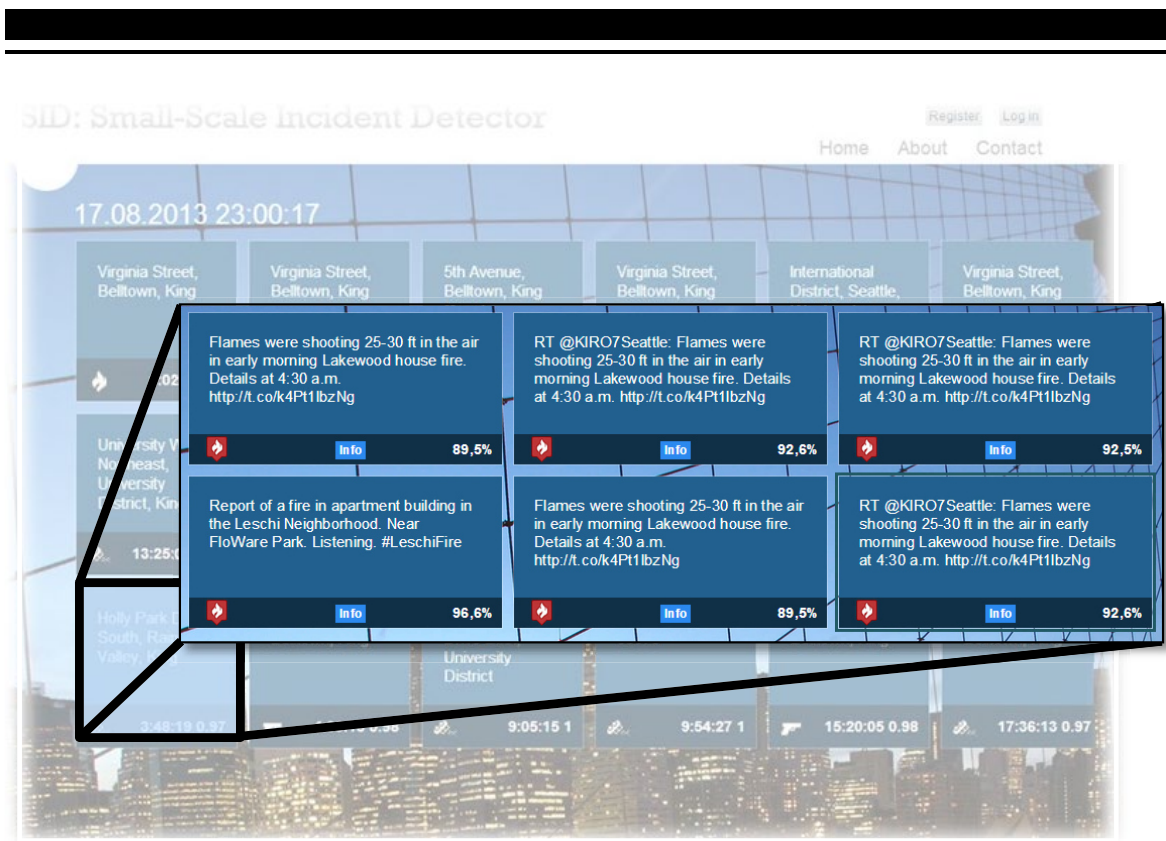


Figure 44.: SID as an application for incident detection based on user-generated content.

8.4 Evaluation

In this section, we present the results of two studies. In the first study, we analyzed which type of *situational information* is contributed by incident-related tweets. This is important to understand (1) if tweets contribute additional information about small-scale incidents and (2) if this information is previously unknown to the decision maker. For this study, we made use of the mechanisms developed throughout this dissertation and conducted a quantitative and qualitative analysis of incident-related information. As a result, we show that valuable information is indeed shared for small-scale incidents. In the second study, we present the evaluation of our approach for detecting incidents based on user-generated content. The goal of that evaluation was to proof the applicability of our approach for incident detection.

8.4.1 Study 1: Analysis of Incident Reports

In this section, we present the results of three studies of incident-related tweets (i.e., of incident reports):

-
- First, we were interested to get an understanding of who is contributing incident reports (i.e., the different types of users). This information is important as some sources provide content that is more valuable compared with others' [105]. Furthermore, as most of the tweets could be sent automatically by emergency management systems, the information would not be new; thus, not valuable. For this, we present a study for analyzing the different types of users posting information about incidents on Twitter.
 - Second, as we are interested in information contributing to situational awareness, we present the results of a quantitative analysis of incident reports. For this, we analyzed the usage of location mentions, temporal expressions, and URLs in incident reports.
 - Third, a qualitative content analysis of incident reports was conducted. In contrast to the quantitative study that focused on the mere presence of situational information, this study was conducted as an in-depth analysis of the information shared.

We first introduce the data set used for all studies followed by the approach and evaluation results of each individual study.

8.4.1.1 Data Set

As there are no public data sets available for qualitative and quantitative analysis of incident reports, we needed to create our own data set. As an initial data set, we used SET_CITY_1, which was collected using the Twitter Search API (see Section 3.2.1) for two North American cities Seattle, Washington, and Memphis, Tennessee. As we needed a high-quality ground truth for our experiments, our initial data sets needed to be further reduced. Thus, we applied the incident keyword filtering presented in Section 3.2.2. From the resulting set, we randomly selected subsets of tweets from the resulting sets containing at least one incident-related keyword. Based on the filtered SET_CITY_1, we randomly selected 1,200 tweets. Those tweets were manually labeled according to our three incident types by five researches of our departments who have experience in emergency management and data labeling. Every tweet was labeled by each researcher. To assign the final coding, four out of five coders had to agree on a label. If no agreement could be achieved, the final label was resolved in a group discussion. This resulted in a data set containing 656 incident reports:

SET_1_L 213 car incidents, 212 fire incidents, 231 shooting incidents, and 544 tweets related to no incident or another type of incident.

As we were also interested in a qualitative coding of each tweet, we also allowed the annotators to assign additional tags to SET_1_L. For this, each incident report could

be annotated in the survey with free tags of at most three word lengths, describing the content of the tweet. For instance, the following tweet was annotated with "1 killed", "1 injured", "crash", and "interstate".

"1 killed, 1 injured in South Memphis crash on I-240: One person was killed Monday morning in a crash on Interstate..."

Reviewers were told not to assign tags to tweets just stating the existence of an incident as this is expressed with the label assigned. Overall, 1,299 tags were assigned to the incident reports.

8.4.1.2 Study 1.1: Exploration of User Types

In the first part of the study, we show which types of users are contributing incident reports. This information is important as some sources provide content that is more valuable compared with others' [105]. Furthermore, as most of the tweets could be sent automatically by emergency management systems, the information would not be new, thus not valuable.

Methodology

To analyze which type of users are contributing information about small-scale incidents, we used the manually labeled data set SET_1_L. Using the description of the users' Twitter profiles, the two lead authors manually coded all users into different categories. Following the approach described in Choudhury et al. [59], we identified five user categories. Official organizations such as the Seattle Fire Department are categorized as *emergency management organizations* (EMO). Organizations not related to emergencies, such as magazines, are clustered as *other organizations* (ORG). Furthermore, we found specialized traffic reporters or journalists, which are represented as *journalists/bloggers focused on emergency management* (EMJ), in contrast to *other journalists/bloggers* (JOU). Users not present in the other groups are categorized as *individual users* (I).

Results

We were able to identify 246 unique users sharing incident reports. The first bar of each stacked cluster in Figure 45 shows the distribution of the number of users for each category according to the different types of incidents. We can notice that a variety of different individual users (196) are reporting about the three incident types. This finding is important as it shows that many different people share incident-related information. On the other side, only few emergency management organizations (11) and focused journalists (2) are publishing incident reports, which is not surprising as their number is limited for a city.

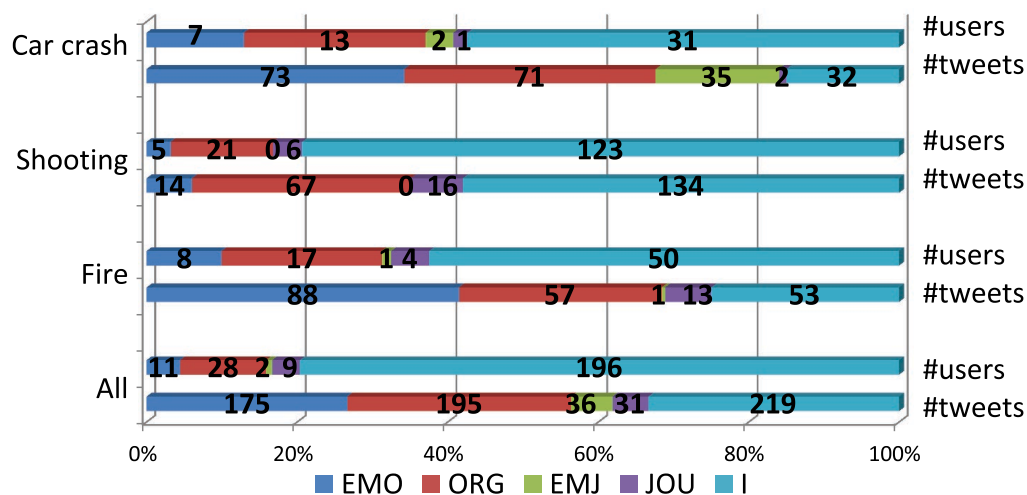


Figure 45.: Distribution of the number of users and the number of tweets from different user categories by incident type in SET_1_L.

The second bar of each stacked cluster in Figure 45 shows the overall number of tweets shared by each user category for the different types of incidents. We can notice that even though the number of the emergency management organizations and other organizations is significantly smaller than the number of individual users, most of the tweets are shared from the users of these organizations (overall 56%). The individual users share 33.3% of the tweets, although the number of tweets by individual users regarding the shootings is much higher. The reason for this might be that shootings are more of public interest compared with car crashes and fires. Furthermore, the results indicate that individual users contribute only one or at most two tweets regarding small-scale incidents.

8.4.1.3 Study 1.2: Quantitative Analysis of Situational Information in Incident Reports

In this part of the study, we show the results of analyzing the different situational features in tweets based on a quantitative analysis of SET_1_L. In this study, we applied automatic techniques for counting the presence of URLs, location mentions, and temporal expressions.

Methodology

For identifying URLs, location mentions, and temporal expressions, we had to apply three different approaches. First, the detection of URLs in tweets was done using regular expressions. For identifying location mentions and temporal expressions, we reused the mechanisms described in Chapter 4 for automatic coding. It is impor-

tant to note that we did not analyze the metadata of a tweet, but the information provided in the message itself.

Results

The automatic data coding allows us to examine the differences in terms of the characteristics of content posted by different user categories. We automatically counted the number of spatial and temporal mentions in the tweets as they provide information about the location and the time of an incident. Furthermore, as URLs are often posted as references to describe pictures or additional descriptions, we analyzed the numbers of URLs based on regular expressions.

Table 52.: Content characteristics of tweets differentiated by user type.

Car Crash				Fire			
Type	Location	Time	URL	Type	Location	Time	URL
EMO	98.63%	15.07%	4.11%	EMO	97.73%	38.64%	11.36%
ORG	91.55%	12.68%	56.34%	ORG	89.47%	40.35%	100.00%
EMJ	100.00%	11.43%	0.00%	EMJ	100.00%	0.00%	0.00%
JOU	100.00%	50.00%	100.00%	JOU	84.62%	38.46%	84.62%
I	59.38%	15.63%	18.75%	I	49.06%	39.62%	37.74%

Shooting				All Incident Types			
Type	Location	Time	URL	Type	Location	Time	URL
EMO	92.86%	28.57%	42.86%	EMO	97.71%	28.00%	10.86%
ORG	82.09%	26.87%	97.01%	ORG	87.69%	25.64%	83.08%
EMJ	0.00%	0.00%	0.00%	EMJ	100.00%	11.11%	0.00%
JOU	87.50%	0.00%	37.50%	JOU	87.10%	19.35%	61.29%
I	62.69%	22.39%	36.57%	I	58.90%	25.57%	34.25%

In Table 52, the results of our analysis are shown. We can conclude that at least for our data set, organizations and journalists tend to always mention spatial locations, while only around half of the tweets shared by individual users contain location mentions. As location mentions can be on country, city, or even street level, we present results on the level of detail of the mentioned location information in the following qualitative analysis. Regarding temporal mentions, no clear differences between the user types can be found. Most of the temporal mentions are shared during fires or shootings compared with less mentions during car incidents.

Most of the links are posted by organizations, journalists, and citizens. In contrast, users part of the EMO category usually do not include URLs in the tweet, which is probably because they tweet early about the incident and there is still no web content that can be published yet.

Finally, we compared the differences in terms of characteristics between incident reports and tweets not related to incidents. As shown in Table 53, incident reports contain twice as much location mentions compared with regular tweets. This gives us a clear indication that spatial and temporal filtering is indeed applicable for incident reports.

Table 53.: Content characteristics of incident reports and tweets not related to incidents.

All Incident Types			
Type	Location	Temporal	URL
Incident Related	81.40%	25.15%	41.92%
Not Incident Related	43.84%	19.18%	43.84%

Regarding the temporal mentions as well as the URLs, we could not find clear differences between incident and not incident-related tweets. Nevertheless, the amount of tweets with temporal mentions is quite high. The results show that during small-scale incidents lots of valuable situational information is shared. In most cases, spatial information referring to the situation where an incident occurred is posted.

8.4.1.4 Study 1.3: Qualitative Analysis of Situational Information in Incident Reports

Following the previous study, we present the results of a qualitative analysis of situational features shared in tweets based on SET_1_L. For example, as we showed that location mentions are commonly present, we wanted to find out how precise location information in tweets is. Furthermore, we show that several other important situational updates are shared.

Methodology

For a qualitative analysis, we identified and organized situational information into categories following the approach of [236]. Based on our data set, we were able to identify situational information of five different categories:

- *Precise Incident Type* is a more fine-grained description of the incident type.
- *Affected Objects* refers to affected objects such as buildings or cars that were damaged.
- *Damage/Injury Reports* are information describing the condition of involved people.
- *Road Conditions* is a description of the road conditions.

-
- *Precise Location* is a description of the location on street level.

The categories were identified based on the qualitative coding, which was conducted on SET_1_L. Each type of information that appeared more than five times was given a category name. Each tweet may be coded with no, one, or more than one category. For example, the following tweet provides information about possible injuries, road conditions, as well as precise location information:

*Traffic: Still dealing w/ MAJOR delays *BOTH* directions on I-240 (Mid-town) near S Pkwy due to early injury crash! #WREG #MEMtraffic*

Finally, the two lead authors assigned the categories to all 656 incident reports based on the tags provided in the survey.

Results

In Table 54, the percentages for each user type and each category are shown as well as the overall percentage of the appearance of each category in all incident reports. Overall, around 10% of all incident reports contain information about the precise incident type, which might be helpful for the fine-grained differentiation of the situation at hand. Most of those tweets are posted by organizations compared with rather a low percentage by individual users.

Information about affected objects is shared quite often in incident reports. Most of those tweets are contributed by ORGs and individual users. As it is highly important for emergency managers to know if a school, a chemistry plant, or a truck carrying flammable liquids is on fire, the early reporting by individual users along with this information can be very helpful. Also, around 21% incident reports contain information about the people involved and the number of injured persons.

Seven percent of the information is related to road conditions. Furthermore, this type of information is mostly shared by official organizations or EMJs. On the other side, precise location information, which is mostly accurate on street and intersection levels, is shared in 20% of the tweets. Also, one-third of this information is provided by individual users.

Furthermore, important situational features such as precise location information, information about the type of event, affected objects, and injured people is shared in incident reports. Thus, good means are necessary to make use of this source of information.

Table 54.: Percentage of situational feature categories for each user type per category and in relation to the overall amount of tweets per user type (in parentheses).

Precise Incident Type		Affected Objects		Damage/Injury reports	
All	9.58%	All	21.46%	All	21.16%
EMO	23.80% (10.0%)	EMO	7.80% (7.33%)	EMO	10.07% (9.33%)
ORG	52.38% (17.93%)	ORG	37.59% (28.80%)	ORG	48.92% (36.96%)
EMJ	1.58% (2.94%)	EMJ	2.13% (8.82%)	EMJ	0.00% (0.00%)
JOU	4.76% (11.11%)	JOU	10.64% (55.56%)	JOU	2.88% (14.81%)
I	17.46% (4.98%)	I	41.84% (26.70%)	I	38.13% (23.98%)

Road Conditions		Precise Location	
All	7.31%	All	19.93%
EMO	27.08% (8.67%)	EMO	19.08% (16.67%)
ORG	43.75% (11.41%)	ORG	32.82% (23.37%)
EMJ	25.00% (35.29%)	EMJ	9.92% (38.24%)
JOU	0.00% (0.00%)	JOU	6.87% (33.33%)
I	4.17% (0.90%)	I	31.30% (18.55%)

8.4.1.5 Study 1: Summary

In this study, we conducted an in-depth analysis of situational information shared in incident-related tweets posted in two North American cities. In the first part of the study, we showed that a variety of individuals are sharing information about small-scale incidents. This is an important finding as this information is not necessarily available for decision makers in the command staff. Furthermore, this also underlines that automatic approaches might have to deal with a large diversity of incoming information such as different styles of writing.

Our manual analysis from a quantitative as well a qualitative point of view showed that these tweets might be helpful as they provide important situational features. First, precise location information is present, which enables decision makers to easily geolocalize the location of an incident. As only around 1% of all tweets are explicitly geotagged, extracting this spatial information from the tweet message is helpful. Second, affected objects such as buildings or cars and much more important information about potentially injured persons are shared. This information is especially valuable as it allows the better planning of response measures. Finally, information about road conditions is shared as well. This information can also be beneficial for citizens alike as it could be used for real-time traffic avoidance.

8.4.2 Study 2: Evaluation of Incident Detection

In this section, we present the evaluation of our approach for detecting incidents based on user-generated content. For this evaluation, several approaches presented in previous sections are used: our approach for detecting the date and time of an event (see Section 4.2), the approach for the geolocalization of tweets (see Section 5.3), the machine-based classification for inferring the type of incident (see Section 7.3), and the event clustering presented in this section.

The goal of this evaluation was to proof the applicability of our approach for incident detection. As shown in the related work section (see Section 8.1), no official data sets are available for evaluating incident detection based on user-generated content. Thus, we decided to use official Open Governmental Data about incidents as ground truth data, which allows easy comparison. Though incident clusters are created, the evaluation did not deal with the aggregation quality of incident reports in each cluster as our main goal is small-scale incident detection. Furthermore, the matching across vocabularies was also not evaluated.

In the following, we first present the experimental setup. Next, we present the results of two studies. The goal of the first study was to evaluate our approach in comparison with other approaches using the real-world incident information. The second study aimed at evaluating the precision of our incident detection in general.

8.4.2.1 Study 2: Evaluation Design, Method, and Procedure

Data Set and Aggregation: As one of our goals was to evaluate the performance of incident detection, we decided to gather information from existing emergency management systems. For this, we used the "Seattle Real Time Fire 911 Calls" data set⁵⁵, which is official emergency information that is provided shortly after an incident is reported to the Seattle Fire Department. Though the data set provides a huge amount of incident information per day, we do not expect it to be a complete set about all real-world incidents in Seattle. However, as this is an official source, it provides high-quality ground truth data about small-scale incidents.

For comparing the real-world incidents with incidents determined from tweets, we collected a two-day sample of tweets using the Search API in a 15 km radius around Seattle, Washington. For this study, we used the whole and unfiltered data set to show the applicability of our approach on a large amount of data. In this case, we did *not* apply any further prefiltering of the data set. This resulted in the following data set:

⁵⁵ <https://data.seattle.gov/> [Accessed: 01.02.2014]

-
- **INC-TW:** 802K tweets for Seattle from March, 11, 2014, 00:00:00 to March, 13, 2014, 00:00:00

Furthermore, we collected real-world incident data in the same period to allow the correlation of tweets and real-world incidents. As the incident types provided by the Seattle Fire Department are more fine-grained than our three types, we aggregated the incident types present in this data set to match our types (see Table 55 for an overview). The data set used for our evaluation consists of 84 real-world incidents from Seattle:

- **INC-S:** 21 car incidents, 61 fire incidents, and 2 shooting incidents

Algorithms: In the evaluation, we compared different state-of-the-art clustering approaches with our approach. As we aimed at detecting small-scale incidents, we wanted to detect real-world incidents within a 500 m radius and within a 20 min time interval. These parameters were provided by emergency managers; thus, all approaches were restricted accordingly.

For matching real-world incidents to information published in tweets, we needed to aggregate all tweets describing the same incident. For comparing with state-of-the-art approaches, we followed the strategies of related approaches and applied a thematic clustering of the tweet data set. For this, we reimplemented two clustering approaches for comparison that do not need the number of clusters to create as input. For comparison, we reimplemented a density-based clustering using DBScan [69] as used by [31]. The approach is dependent on two parameters "Epsilon" and "Min points", which define the proximity and density of instances. As the optimal values are not known in advance, we evaluated several combinations of both parameters. As all resulting clusters needed to be manually labeled, we could not evaluate all possible combinations but used a greedy approach to heuristically identify those that likely provide the best results.

As a second clustering approach, we reimplemented an online single-pass clustering as used by [173, 22, 108]. We decided to use Leader-follower clustering [193]. In this approach, a threshold needed to be specified to determine if an instance is merged to a cluster. Also for this approach, several parameter settings were evaluated.

For both approaches, we used the Euclidean distance similarity metric based on a tweet's tokens weighted using the TF-IDF scores. Furthermore, as we were interested in a 20 min temporal extent and due to clustering performance, we separated INC-TW into 20 min time intervals and applied the respective algorithms on window. To finally allow correlation with real-world incidents, we conducted a spatio-temporal filtering. We selected only clusters that are within a 500 m radius and a ± 10 min time interval of a real-world incident. For applying the filtering, we followed the same spatial and temporal merging strategy as used in our approach (see Section 8.2).

Thus, spatial locations of the individual tweets in a cluster that have GPS coordinates are merged to a weighted mean location. As a temporal dimension for a cluster, the creation date of the oldest tweet in the cluster was used. In contrast to related approaches, we did not apply a supervised classification of the resulting clusters as this would need to be trained with additional data. We decided to manually evaluate every resulting cluster to check whether it corresponds to the real-world incident.

We compared the results of these approaches with those of our approach. For our approach, we applied the spatio-temporal-thematic clustering as described in Section 8.2. The steps conducted for our approach are the following:

1. **Classification:** For detecting the type of incident (i.e., the thematic dimension of each information item), the machine-based classification (see Section 7.3) was applied. For this, a model was trained on data set 4-CLASSES comprising 2,000 tweets. The classifier was then applied to all tweets in the newly collected data set for Seattle, INC-TW. All tweets with probability $> 70\%$ were chosen for further processing. Choosing a lower probability would result in too many misclassified tweets. As a result, we identified 1,271 incident-related tweets with the respective incident types. These tweets were sent by 685 distinct users.
2. **Spatio/Temporal Resolution:** Next, the temporal and spatial dimensions are inferred for each incident-related tweet. For this, our approach for detecting the point in time of an incident (see Section 4.2) and the approach for the geolocalization of tweets (see Section 5.3) were applied on each tweet. This gave us spatio-temporal information for each incident report, which is needed for our rule-based clustering. Most importantly, only 25 out of 1,271 tweets contain an exact GPS coordinate; thus, more than 98% of the tweets would not be used by related approaches.
3. **Rule-Based Clustering:** Finally, the rule-based event clustering was applied on the prefiltered tweet data set containing all the metadata. We created a single rule $\{Incident_Type, 500\ m, 10\ min\}$ for each incident type (Crash, Fire, Shooting) with a very strict temporal boxing. Furthermore, we decided to aggregate all information within 500 m to take the estimation error of the geolocalization approach into account. Using the algorithm, the 1,271 tweets were aggregated to 366 distinct incident clusters.

Metrics: As metrics, we present precision and recall as defined by [136] and [243] for evaluating the accurateness of event detection:

- *Recall* is the percentage of real-world incidents detected of all real-world incidents in the same time period.
- *Precision* is defined as the percentage of incidents detected of all events detected.

Both definitions rely on the assumption that an incident is detected if at least one tweet is contained in an incident cluster matching the real-world incident.

For the first study (i.e., the correlation of incidents derived from the tweets with real-world incidents), we calculated the recall of all detected incidents and the real-world incident information. To decide whether an incident detected with our approach matches the real-world incident, all tweets in each incident cluster are manually compared with respect to the incident type and the spatial location mentioned in each tweet. If at least one incident-related tweet was contained, a match of the incident cluster to the real-world incident was confirmed.

For evaluating the performance of the overall approach in the second study, we calculated the precision of the resulting incident clusters. This time, if at least one incident-related tweet was contained in a cluster, we assumed that we detected an incident.

8.4.2.2 Study 2: Results

The results for evaluating the correlation to real-world incidents are shown in Table 55. In the case the spatial and temporal boxing is applied, 32.14% of the real-world incidents could be detected. This is a very precise result given the strict temporal and spatial boxing. Furthermore, if no spatial boxing was applied, we were able to detect 57.14% of the incidents reported in the official emergency system using only information present in tweets. Also, the individual recalls for each incident type are high with more than 50% without spatial boxing and more than 29% with spatial boxing for each incident type. However, the results for the shooting incidents may not be representative as only two shooting incidents occurred in this time period.

We identified several limitations of our approach, which are open for future investigations. First, for some events, no information could be found. This might be a reason of the incomplete sample provided by the Twitter Search API. Second, some tweets were classified wrong, which is likely a reason of the rather small data set used for training. Third, the spatial bounding box reduces the overall detection rate; thus, a more fine-grained approach for geolocalization is needed. Nevertheless, the approach allows detecting one-third of the real-world incidents.

Table 55.: Correlation results of real-world incidents to incidents mentioned in tweets.

Incident Type	Real-World Incident Type	# of Real-World Incidents	Recall	Recall (<500 m)
Car Incident	Motor Vehicle Accident	21	71.42%	38.09%
	Motor Vehicle Accident Freeway			
	Medic Response Freeway			
	Car Fire Car Fire Freeway			
Fire Incident	Fire In Building	61	52.46%	29.50%
	Fire In Single Family Residence			
	Automatic Fire Alarm Residence Auto Fire Alarm			
Shooting Incident	Assault w/Weapons per Rule Assault w/Weapons Aid	2	50%	50%
All		84	57.14%	32.14%

In Table 56, a comparison of our approach with related approaches is shown. The table shows the best results that could be retrieved with the respective parameter settings. A complete overview of all results is shown in Table 69 and Table 68 (see Appendix A).

Table 56.: Comparison of approaches for small-scale incident detection.

	Recall (<500m)
Our Approach	32.14%
Single-pass clustering (Leader-follower, Threshold=0.8)	5.88%
Density-based clustering (DBScan, MinPoints=1, Epsilon=0.4)	3.53%

The results indicate that we are able to detect five times more real-world incident compared with the best related approach, which has a recall of 5.88%. One reason for the low recall of related approaches is that only tweets with GPS coordinates are used. Thus, the overall number of clusters with GPS information is of course rather low. However, as this is the current state-of-the-art approach for event detection, the results are comparable to our approach.

Another disadvantage of related approaches is that the clusters that are created are not yet labeled as incident clusters or clusters with information not related to incidents. For the evaluation, these clusters were labeled manually, providing accurate results, whereas for real-world applicability, an additional approach is needed for differentiating relevant from irrelevant clusters. The results show that applying clustering techniques as used in related work such as single-pass incremental clustering results in a set of clusters containing many tweets. For instance, with the best parameter settings for DBScan, eight clusters were detected with an average number of 478 tweets. With the leader-follower approach, 132 clusters were created with about an average number of 5 tweets. In this set of clusters, many irrelevant clusters are contained, which would need to be excluded with high computational costs. In contrast, with our approach, we exclude potentially irrelevant information before clustering, giving us only a set of incident clusters. Furthermore, the clusters created with our approach are small in size, with an average of four tweets; thus, detecting relevant information is much easier done manually.

In a second evaluation, we manually analyzed the aggregated tweets for 100 incident clusters created with our approach. We found that for 77 clusters, at least one incident-related tweet was contained, whereas the other 23 clusters did not contain incident-related information. This gives us a precision of 77% for detecting incidents based on tweets. Though other approaches for small-scale incident detection are not directly comparable as a different data set was used, we provide results comparable with [5] that reported a precision of 76.6% and [103] that reported a precision of 72%. Though the precision is comparable, none of the related approaches provide a spatial and temporal accurateness as we do.

Furthermore, the precision of 77% is quite surprising as this incident information was not contained in the official emergency management system. A reason for this could be that some incidents are not covered by the official system, for instance, if no police was called. However, the results allow detecting many potentially valuable information for emergency management.

As another important result, we found that an average of four tweets is contained in each incident cluster. This underlines that detecting small-scale incidents with such a low amount of data is possible with our approach. Furthermore, the average time of each incident cluster is three minutes and 52 seconds after the real-world incident was reported to the emergency management organization. This shows that indeed timely information can be retrieved.

8.4.2.3 Study 2: Summary

The evaluation results show that we outperform related approaches for detecting events such as real-world incidents. In the evaluation, we showed that the combination of the different approaches presented throughout this dissertation in addition to our approach of rule-based event clustering is capable of detecting more than 50% of real-world incidents published in an emergency management system. Furthermore, we detect 32.14% of the incidents within a 500 m radius and within ± 10 min. These results are more than five times better compared with related approaches. This is because we provide a much higher number of geotagged information as well as more accurate temporal information, allowing us to create better incident clusters. Furthermore, the overall number of truly incident-related clusters is much higher with our approach as we filter out not incident-related information before clustering. We also showed that even more incident-related information can be found in tweets, which we are able to detect and cluster precisely.

8.5 Conclusion

In this section, we presented an approach for incident detection and clustering of incident reports. Specifically, we showed how to automatically detect incidents based on spatial, temporal, and thematic information. As a result of this approach, clusters of information related to the same incident are created. The presented approach also addresses the problem of different vocabularies of incident types used across emergency management organizations.

In this chapter, we made the following contributions:

- We proposed a spatio-temporal-thematic clustering algorithm, which is able to detect incidents in a large amount of social media data. Furthermore, the ap-

proach clusters all information related to the same incident. In contrast to previous approaches, we allow precise spatio-temporal localization. Furthermore, our clustering approach is able to deal with different incident type vocabularies. This is important to overcome communications issues that derive from different vocabularies.

- In a study, we conducted an in-depth analysis of situational information shared in incident-related tweets posted in two North American cities. We showed that a variety of individuals are sharing information about small-scale incidents not necessarily available for decision makers in the command staff. Furthermore, a qualitative and quantitative analysis of incident reports showed that these tweets contain indeed valuable information for improving situational awareness. The results underline the value that user-generated content could contribute for decision making in the emergency management domain.
- In an evaluation of our clustering approach, we showed that it is capable of detecting more than 50% of real-world incidents published in an emergency management system. Furthermore, we detect 32.14% of the incidents within a 500 m radius and within ± 10 min. These results are more than five times better compared with related approaches. This is because we provide a much higher number of geotagged information as well as more accurate temporal information, allowing us to create better incident clusters.

Furthermore, we showed that more than 77% of the detected events are related to incidents. In contrast to related works, the overall number of truly incident-related clusters is much higher with our approach as we filter out not incident-related information before clustering. As a result of our approach, decision makers are now able to consume previously completely unstructured user-generated content in such a way that it is possible to take it into account for decision making in emergency management.

For future work, several known limitations of our approach can be addressed. As shown in the evaluation, different sources of user-generated content might be integrated into the clustering process to allow the detection of more incidents. Furthermore, the approach for the geolocalization of tweets as well as the machine-based classification can further be enhanced with more training data to achieve better results.

Also, important information such as the number of injured people or affected buildings already present in the aggregated event clusters could be detected automatically. Furthermore, applying multi-label classification in combination with different organization-specific vocabularies is an aspect that needs further research.

In the next chapter, we show how to adapt the overall framework to deal with the dynamism of social media platforms.

9 Refinement of the Framework for Detecting and Clustering Incident Information

In the last chapters, we presented our approach for detecting incidents based on user-generated content as well as the approach for clustering related incident reports. In this chapter, we show that active learning can be used for reducing the amount of instances that need to be annotated (or labeled) for refining the pipeline (see Figure 46). As a result of this step, the machine-based classification step can be refined with new labeled instances.

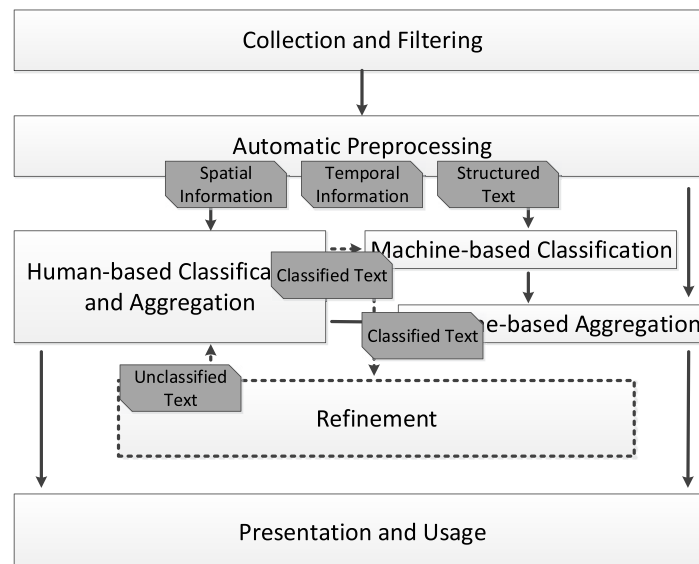


Figure 46.: Refinement as a step in the framework.

We already outlined that social media platforms such as Twitter are not static environments; thus, classification of new data is necessary so the pipeline can be refined according to the changing conditions. Also, adapting the current pipeline to other incident types makes labeling of new data a necessity. In Chapter 6, we showed how crowdsourcing can be facilitated to classify new data. However, crowdsourcing only scales for large events, and it has limited applicability when real-time information is needed. Furthermore, the more annotations are needed, the more expensive a crowdsourcing campaign will be. To solve the challenge of high labeling costs and timely information retrieval, the size of the data set that is to be labeled needs to be reduced while maintaining highly accurate classifiers.

To this end, *active learning* has been proposed [208]. Active learning aims to reduce labeling costs by iteratively (1) selecting small subsets of instances (e.g., tweets) to query for annotations and (2) retraining a classifier until the classification quality is sufficient. The subset to annotate is selected such that the classifier achieves the highest accuracy with as few labeled instances as possible. However, it is still an open research issue how to maintain the most accurate model, especially when it comes to applying active learning on user-generated content. Therefore, we deal with the two following fundamental research questions in active learning:

- **Q1:** How can the initial training set to train the initial classifier be selected?
- **Q2:** How can appropriate instances for labeling in each iteration be chosen?

The selection of the initial training set is of great importance for the starting performance of the classifier. Though the classifier learned on a suboptimal selection can still be improved in further iterations, starting off with a high-quality initial training set keeps the number of iterations low and guarantees that the active learning procedure is not hindered by a bad start. Furthermore, selecting good instances during each iteration ensures a fast increase of classification performance.

Also, active learning approaches usually assume perfect annotation quality (= a perfect oracle) [208]. As we showed in Section 6.5.2, this is not true for crowdsourced annotations as they are prone to errors. Thus, we need to answer the additional question:

- **Q3:** How do errors influence the learning process?

In the following section, we present a clustering approach that leverages temporal, spatial, and thematic metadata for the selection of the initial training set (Q1) as well as the query instances (Q2). We show that our approach yields a decreased deficiency compared with state-of-the-art approaches. Furthermore, the approach is more robust to labeling errors compared with related approaches. Our evaluation also shows that in contrast to a plain supervised approach, less training instances are needed; thus, reducing the overall costs.

The contributions presented in this chapter are the following:

- We present two algorithms on the basis of event-based clustering for identifying the initial training set and the training sets for each iteration. Compared with other approaches, we make use of spatial, temporal, and thematic information.
- We validate the effectiveness of our approach on a data set of incident-related tweets compared with state-of-the-art approaches. We show that our approach outperforms related work.
- We validate the influence of noise on our approach using quantified error rates.

- With the presented approach, labeling costs can be reduced as fewer training data need to be labeled to achieve similar F1 scores compared with a supervised learner that has access to all training data.

In the next section (see Section 9.1), we give an overview of active learning and how to make use of active learning for the classification of user-generated content. Following, we give an overview of related approaches in Section 9.2. Next, we describe our active learning approach in Section 9.3 and its evaluation in Section 9.4. Finally, we close with a conclusion and future work in Section 9.5.

9.1 Background

In this section, we provide an overview of active learning and how to make use of it for the classification of user-generated content. We decided to use active learning as in contrast to other semi-supervised learning approaches, it aims to minimize annotation effort [208].

As introduced before, active learning can be used in an iterative process to build classification models by selecting only subsets of the available instances to annotate. In Figure 47, a high-level overview of the active learning steps integrated into our framework is shown. The steps consist of two main parts: (1) a learning step, where a classifier is built, and (2) a refinement step, in which the classifier is retrained with new labeled data.

In this dissertation, we assume that we have a large collection of unlabeled data available at once. Thus, we follow a pool-based sampling approach [134] in which the whole collection is used to decide which instances to label next.

As a starting point of our framework, user-generated content is collected as an initial pool of unlabeled data U . From this information base, a set of training examples L has to be chosen for learning an initial model. This problem is called the *initial selection problem* (corresponds to Q1). It is highly important how to choose this set because with a well-selected initial training set, the learner can reach higher performance faster with fewer queries [116]. In most cases, this selection is done randomly [256], which cannot guarantee an appropriate selection of instances. Based on this initially labeled set, training a classifier is performed.

After training, this classifier is refined in several iterations in the refinement step. In each iteration, b instances are selected for labeling. The labeled instances are removed from the pool of unlabeled instances U and added to the pool of labeled instances L ; thus, more instances can be used for learning. For selecting these instances, a *selection strategy* is used on U to query labels for a number of instances in each iteration. This is known as the *query selection problem* (corresponds to Q2).

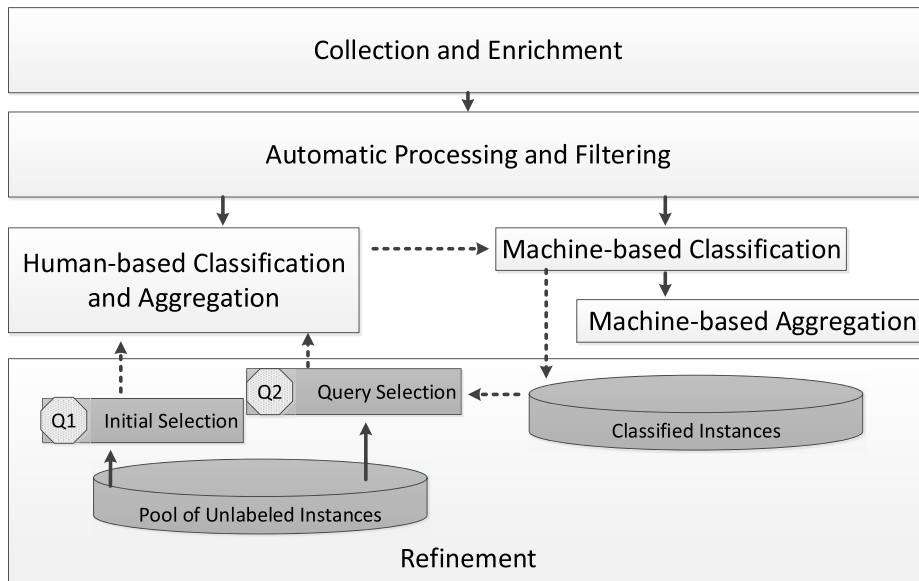


Figure 47.: Initial selection and query selection as parts of the refinement step.

For solving both problems, several selection strategies can be chosen based on two criteria, informativeness and representativeness [104]:

- *Informativeness* measures the usefulness of an instance to reduce the uncertainty of the model.
- *Representativeness* measures how good an instance represents the overall input of unlabeled data.

In Figure 48, an example of a binary classification problem is given with respective decision boundaries of an SVM. The figure shows these instances that would be selected with the respective selection strategies.

For informativeness as selection criteria, the two most prominent approaches are uncertainty sampling and query by committee (QBC). In uncertainty sampling, different measures might be applied. For instance, a least confident approach would consider the instance for which the prediction is least confident as uncertain, whereas the entropy-based uncertainty sampling calculates the uncertainty based on the entropy measure [210]. In the case an SVM is used as a classifier and uncertainty sampling is applied, instances closest to the decision boundary are chosen [208] (see Figure 48).

In contrast, QBC is a multi-classifier approach in which the disagreement between different classifiers is used to determine the informativeness of an instance [166].

The main issue with the informativeness approach is that only a single instance is regarded [208]; thus, outliers could be selected erroneously as the context is not taken into account (see Figure 48, a)). Furthermore, the selection of query instances may be determined by a small number of labeled instances [104].

In the case of identifying representative instances, the structure of the data is taken into account [56]. Thus, several clustering strategies are employed (see Section 9.2). According to Nguyen and Smeulders [162], the most representative examples are those in the center of the cluster, which are the instances most similar to all other instances in the cluster. Nevertheless, the performance of query strategies based on this approach is dependent on the clustering results and how instances are chosen from these cluster (see Figure 48, b)). For instance, always selecting the centers of the clusters might result in always selecting very similar instances for each iteration; thus, the model might not improve very much. Furthermore, it remains unclear how many clusters have to be built. Also, and most important in our case, the resulting clusters do not necessarily correlate to real-world events as spatial and temporal information is omitted.

To overcome the individual problems of each approach, related work proposes to combine both (see Section 9.2). This results in selecting the instances that are representative for the whole data set as well as have the highest chance to improve the model. In our approach, we use metadata provided in microposts to cluster instances based on both criteria and to choose the most valuable instances for training the classifier (see Section 9.3). The whole process of refinement and classification continues until some stopping criteria such as a maximum number of iterations is reached or when the model does not improve anymore.

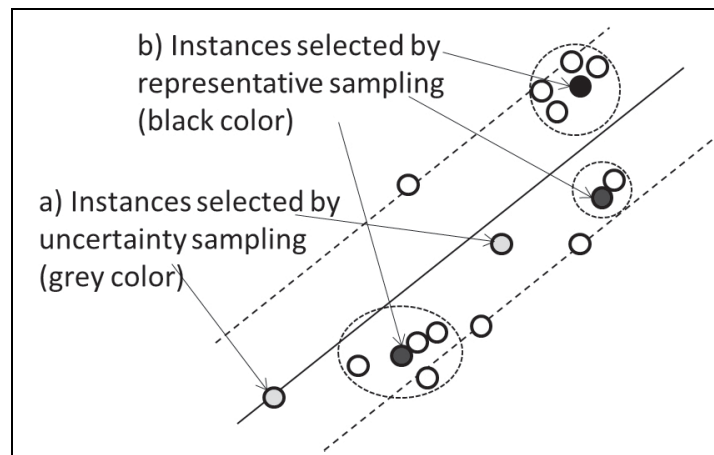


Figure 48.: Example of a binary classification problem with respective decision boundaries of an SVM. Furthermore, the instances selected by uncertainty sampling and representative sampling are shown.

9.2 Related Work

In this section, we present related approaches with respect to three dimensions: First, we give an overview of approaches that apply active learning on tweets. Second, we analyze selection strategies used in related works. Third, we show how noisy annotations are taken into account in the active learning process. In Table 57, an overview of these related approaches is given:

Table 57.: Comparison of related active learning approaches with respect to selection strategies and the use of event-related metadata.

Approach	Social Media	Metadata	Selection Criteria	
			Informativeness	Representativeness
Thongsuk et al. [229]	x		x	
Hu et al. [102]	x			x
Seung et al. [209]			x	
Freund et al. [76]			x	
Balcan et al. [18]			x	
Lewis and Catlett [133]			x	
Tong and Koller [231]			x	
Xu et al. [249]				x
Kang et al. [116]				x
Zhu et al. [256]				x
Nguyen and Smeulders [162]				x
Donmez et al. [63]				x
Tang et al. [225]			x	x
Shen et al. [211]			x	x
Donmez et al. [63]			x	x
Zhu et al. [256]			x	x
Huang et al. [104]			x	x
Our Approach	x	x	x	x

9.2.1 Active Learning on Social Media Data

Although active learning is a mature field [50] and has been studied extensively for text classification [99, 231, 151, 134], active learning on short texts such as tweets was only proposed by two previous works. Thongsuk et al. [229] presented a technique for classifying Twitter posts into three business types. They showed that using active learning outperforms simple supervised learning approaches in terms

of costs. The used approach is based on entropy-based uncertainty sampling. It remains unclear how they built the initial set.

Hu et al. [102] presented the ActNeT approach, which takes the relations between tweets into account for identifying representative and informative instances. Based on the social network, the topology is used to detect representative instances using the PageRank [168] algorithm. Informative instances are chosen using an entropy-based uncertainty sampling. However, as related tweets are not necessarily available and have to be collected with great effort, their approach is not directly usable in the real world. Furthermore, tweets whose author has no friends are omitted even though they could contain valuable information.

Though both approaches are specifically designed for tweets, both omit noisy annotations. Furthermore, none of the approaches take additional metadata about an event into account.

9.2.2 Selection Strategies for Active Learning

Selection strategies based on informativeness are the most popular approaches for active learning. For instance, QBC was used by [209, 76], whereas uncertainty samplings was used by [18], [133], [231].

Most related to our approach are clustering strategies, which are employed for identifying the most representative instances. In this area, different cluster-based strategies have been proposed for active learning, for instance, based on k -means clustering [249, 116, 256] or k -medoids [162, 63]. The main weakness of these approaches is that the performance highly depends on the quality of the clustering [56]. These approaches might not be suitable for incident detection, as the number of clusters needs to be specified beforehand, which is unknown in our case. Determining the optimal number of clusters would result in high computational costs and would not guarantee an optimal result.

Several approaches were presented to combine the informativeness and representative criteria. Tang et al. [225] used k -means clustering and proposed to select the most uncertain instance for each cluster. Information density was then used to weight instances, unlike to our approach, which uses density in a cluster to select instances. Shen et al. [211] applied k -means clustering and uncertainty sampling. In this case, information density is calculated within a cluster. Multiple criteria are linearly combined with different application-specific coefficients, which are difficult to determine automatically [256]. Donmez, Carbonell, and Bennett [63] proposed to combine uncertainty sampling and k -medoid to identify representative and informative instances. They showed that this approach is beneficial compared with using only a single approach. However, their approach needs to manually set the number of clusters. As these approaches rely on k -means or k -medoid clustering, it is diffi-

cult to estimate the optimal number of clusters. Thus, the average size of the cluster might be very large, which scatters the density calculations as examples might be very close to each other.

The approach of Zhu et al. [256] is the most promising in this area, and we use this approach as a foundation for our technique. Their method uses clustering for the initial selection. Uncertainty sampling is combined with estimating a density for query selection. Unlike their work, we also apply our event-based clustering for the iterations. Huang et al. [104] followed a similar approach. Instances are selected based on clustering and on confidence in predicting a class label as informativeness measure. Though their approach is quite promising, the authors stated that their approach is restricted to binary classification.

The overview shows that some works tried to combine informativeness and representative and showed promising results. Nevertheless, none of these approaches have been evaluated with noisy labels and on short texts such as tweets.

9.2.3 Active Learning and Noisy Labels

The general assumption of perfect labels in active learning is still open to research [208]. Until 2008, active learning techniques did not tackle the problem of different reliabilities of oracles.

Sheng et al. [212] and Zhao et al. [255] analyzed several heuristics that take labeling uncertainty into account and showed that the repeated relabeling of wrongly labeled tweets could improve label quality and model quality. Though both experiment with equally and consistently noisy annotators, we showed that this assumption does not hold true for user-generated content (see Section 6.5).

Furthermore, proactive learning proposed by Donmez and Carbonell [62] was developed and extended [64, 65] over the years for selecting the most appropriate annotators for every iteration. They presented a decision-theoretic approach for selecting the most reliable oracle. In this case, they take costs of labeling into account. Wallace et al. [237] enhanced this approach and assumed that annotators are able to estimate the quality of their own labeling. Based on this, they proposed to estimate which instances are difficult to label for choosing appropriate annotators. In the most recent work, Ipeirotis et al. [107] proposed repeated labeling strategies for noisy labels. They showed that both the quality of labeled data and the quality of the trained models can be increased with their approach.

9.2.4 Summary

In the following, we discuss related approaches.

Active Learning on Social Media Data: Only few approaches were applied for the classification of event-related short texts such as tweets. Thus, the special properties of social media data that influence the active learning process are not taken into account by related approaches. Furthermore, the approaches applied on social media data both omit noisy annotations. Furthermore, none of the approaches take additional metadata about an event into account.

Metadata Used and Selection Strategies: Related works can clearly be differentiated into approaches taking either informativeness or representativeness into account as selection strategies. Uncertainty sampling and QBC are used as approaches for selecting informative instances. For identifying representative instances, the k -means or k -medoids algorithms were used. However, the major disadvantage of these approaches is that whenever clustering is applied, the number of clusters needs to be specified a priori. This information is not known for incident detection. Furthermore, none of the related clustering approaches make use of metadata such as spatial and temporal information.

Influence of Noise: Finally, the problem of noisy annotators needs to be taken into account when evaluating our approach as this is not well evaluated yet. Though some approaches already dealt with the problem of noisy annotations, no approach quantified the quality of human-based classifications for their experiments; thus, we are the first to use real error rates for comparison with related work.

In contrast, we present two algorithms on the basis of event-based clustering for identifying the initial training set and the training sets for each iteration. Our approach makes use of spatial, temporal, and thematic information. Also, we analyze our approach with respect to error rates based on quantified error rates.

9.3 Approach

In the last sections, we presented the general overview of active learning, and we showed that selection strategies are needed. In this section, we present our active learning-based refinement strategy. We deal with the initial selection (Q1) as well as with the query selection problem (Q2).

As the selection of the initial set and the selection strategy for the iterations highly influences the results, we first want to tackle the question of which approach could be used for the classification of short and unstructured texts such as tweets. As active learning is an evolving field, many strategies were proposed over the last years. Nevertheless, only few leverage metadata as it is provided in event-related data; thus, we want to show how such information can be used for active learning strategies. Most of the recent selection strategies in active learning approaches focus on identifying *either* representative *or* informative instances. Nevertheless, it has been

shown several times that selecting representative *as well as* informative instances could enhance the active learning process [104].

With our approach, we first use metadata such as temporal and spatial information to perform an event-based clustering based on the approach described in Section 8.2. Using this approach, we are able to find a number of event clusters that are likely related to real-world events. Second, based on this clustering, we propose a novel strategy that covers both selection criteria to identify appropriate instances for active learning.

9.3.1 Motivating Example

We first give an example of a two-class classification problem to motivate our strategy (cf. Figure 49). Most of the current approaches utilize uncertainty sampling, where those instances are chosen for labeling for which the classifier is most uncertain. These are the instances that are near the decision boundary. Nevertheless, as shown in Figure 49, a) those instances might not be beneficial for the overall model as they might be outliers.

As opposed to this, applying clustering helps to identify representative examples as shown in Figure 49, b). According to Nguyen and Smeulders [162], the most representative examples are those in the centers of the cluster, which are the instances most similar to all other instances in the cluster. Nevertheless, always selecting the centers of the cluster might result in selecting always very similar instances for each iteration; thus, the model might not improve very well. Furthermore, it remains unclear how many clusters have to be built.

Based on these findings, the general idea is to select the most informative *and* representative instances as shown in Figure 49, c). This results in selecting the instances that are representative for the whole data set as well as have the highest value for improving the model.

As shown before, in the case of identifying representative instances, different clustering strategies are employed. From each cluster, representative instances, which might be determined based on the similarity of instances or just by using the centroid of a cluster, are drawn. However, the performance of strategies based on this approach is dependent on the number of clusters and how instances are chosen from these clusters.

When it comes to classifying the type of an event, we have no a priori knowledge about the number of clusters; thus, we are compelled to use a naive approach such as the number of distinct event types. Nevertheless, this does not ensure a selection of instances to label that result in the best improvement of the classifier. This is because social media data might vary significantly for each event. For instance, one

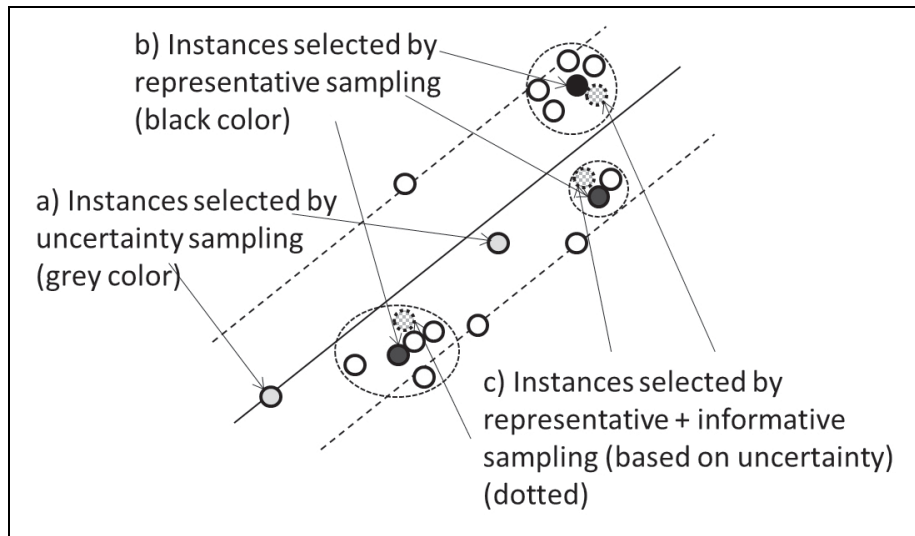


Figure 49.: Instances selected by different selection strategies for a binary classification problem.

event might be a tiny fire in a waste bin, whereas another event is a fire in a factory; though microposts for both events need to be classified with the "fire" event type, training a precise classifier with as few instances as possible is challenging given the classifier has access to instances of only one of these events (because of a suboptimal selection). Thus, for selecting representative instances, it is of great importance to select instances for *both* events. However, current clustering approaches are not directly capable of separating microposts into distinct real-world events; thus, an appropriate selection is difficult. To distinguish these events, a straightforward approach is to create clusters based on the properties of real-world events (i.e., the thematic, temporal, and spatial information). Proceeding this way, the two example events are inherently assigned to different clusters, and hence, instances to be labeled are drawn from both of them. Traditional clustering approaches would rely only on the event type, thus eventually selecting only instances for one event.

Consequently, and in contrast to previous approaches, we contribute a novel event-based clustering approach that also leverages the temporal and spatial dimensions of event-related social media data to allow a more fine-grained clustering. Due to smaller clusters, the selection of appropriate instances is easier because one can assume that even with a bad sampling, the selected instances will still be of high quality. The evaluation shows that this enhanced clustering improves the selection process and that our approach yields a decreased deficiency compared with state-of-the-art approaches. It is also shown that in contrast to a plain supervised approach, fewer instances have to be labeled to reach a comparable performance because in the beginning, there is a very steep increase.

9.3.2 Event-Based Clustering for Active Learning

As shown in the related work section (see Section 9.2), clustering-based approaches are frequently used for identifying representative instances. However, there might not be an obvious clustering of unknown data; thus, clustering might be performed at various levels of granularity. In this case, the number of optimal clusters is unknown and has to be determined with high computation costs. A simple approach might be to use the number of distinct event types as the number of clusters. Nevertheless, this approach is not appropriate as there might be two distinct "fire" events happening in the real-world for which all related information would be integrated into the same cluster. Choosing appropriate instances from this potentially large cluster is difficult. In contrast, we use a more natural way of clustering by taking the properties of real-world events into account. We use event-related information such as temporal and spatial information in combination with the event type to perform an *event-based clustering* as proposed in Chapter 8. Using this approach, we are directly able to find a number of clusters without the need of specifying the number beforehand. Furthermore, our event-based clustering is based on both selection criteria so we overcome the limitations of each individual one.

General Approach for Event-based Clustering

For our event-based clustering, we assume that every event-related information is either related or not related to a specific real-world event. Thus, we propose to cluster all instances based on the three dimensions that define an event: temporal and spatial extent as well as the event type. As a result, each instance is aggregated to a cluster that helps to identify those tweets that might be helpful for better training. Furthermore, compared with other clustering approaches, event-based clustering enables us to calculate the more appropriate number of clusters very easily.

For clustering instances, we use the clustering approach presented in Section 8.2. For this, we make use of thematic as well as temporal and spatial information extracted in the several steps of the overall processing pipeline. All instances are clustered if they lie within the spatial, temporal, and thematic extent of another instance as we assume they provide information about the same event.

Instances containing no thematic information are assigned the *unknown_event* type. Missing spatial information is replaced with a common spatial center (e.g., the center of a city). Missing temporal information is replaced with the creation date of the instance. Thus, even with one or two missing dimensions, we are still able to build a cluster.

Based on this clustering approach, we are able to cluster all instances related to a specific event. This helps to identify those instances that might be helpful for better

training. On the other hand, instances not related to events are contained in larger clusters, containing lots of noise and being less valuable for the learning process.

In the following section, we present the concrete implementations of our approach for the initial selection strategy and the query selection strategy.

9.3.3 Initial Selection Strategy

As a first step, the initial data set that needs to be labeled is selected. As it is difficult to perform the initial selection, related approaches rely on random sampling or clustering techniques [256]. However, these approaches do not guarantee the selection of appropriate instances because the initial sample size is rather small, whereas the size of clusters is large. In contrast, event-based clustering uses the properties of real-world events to perform an initial clustering.

Based on the set of clusters resulting from our event-based clustering, the most representative instances for the complete and unlabeled data set are identified for training the initial model. For this, we use the event clusters ordered by information density of the instances contained in the cluster to obtain a good initial set. Selecting informative instances clearly is not possible yet as a classifier cannot be trained at this point. Our approach for selecting the initial data set is shown in Algorithm 9.

Algorithm 2: Algorithm for initial selection strategy.

Data: Unlabeled instances U , Clusters C generated by event-based clustering,
Size of initial training set b_i

Result: Instances to label L

```

1 forall the clusters  $c$  in  $C$  do
2   forall the instances  $i$  in  $c$  do
3     Calculate information density  $DS(i)$  ;           /* see Eq.23 */
4 forall the clusters  $c$  in  $C$  do
5   Calculate average information density  $DSC(c)$  ;       /* see Eq.24 */
6 Order clusters based on  $DSC$ 
7 while  $|L| \leq b_i$  do
8   forall the clusters  $c$  in  $C$  do
9     Add one instance from  $c$  to  $L$ ;
```

In the following, we describe how the selection of the initial data set is conducted. First, our clustering approach is applied on the complete unlabeled set U without a thematic specification as this is not present at this time. Thus, the *unknown_event* type is used as a thematic extent.

Second, for all instances in each cluster, the information density $DS(x)$ is calculated. In general, the density is calculated based on how many instances are similar or near to each other; thus, outliers are regarded as less valuable. For calculating the density of an instance, we use the k -nearest-neighbor-based density calculation [256]. The density $DS(x)$ of instance x is calculated based on the N most similar instances in the same cluster⁵⁶ $S(x) = \{s_1, s_2, \dots, s_i\}$. As a similarity measure, we use the cosine similarity between two instances. The similarity measure was chosen because it showed good performance for clustering texts (see Section 8.1). For calculating $DS(x)$, the following formula is used:

$$DS(x) = \frac{\sum_{s \in S(x)} \text{similarity}(x, s_i)}{N} \quad (23)$$

The information density DSC of each cluster C is then calculated based on the average of the information density of each instance x contained in C :

$$DSC(C) = \frac{\sum_{x \in C} DS(x)}{N} \quad (24)$$

Doing this, we are able to avoid noisy cluster with lots of unrelated items, which would typically be in clusters not related to an event. Based on $DS(c)$, the clusters are sorted, and instances are selected. For the cluster with the highest information density, exactly one instance is chosen; thus, we avoid drawing too similar instances for the initial training. Proceeding this way, we achieve a good distribution over all valuable event clusters as it is guaranteed that the instances are selected from the most representative clusters. Based on these instances, the initial model is built.

9.3.4 Query Selection Strategy

The initial selection strategy gives us the most valuable instances for training the initial model. For every following iteration, appropriate instances for improving the classifier have to be chosen. As motivated before, our goal for the query selection strategy is to choose representative and informative instances. Besides identifying representative instances based on clustering, the goal of our approach is to avoid

⁵⁶ K is equal to the number of instances in the cluster.

instances that the learner is already confident about. Pseudo-code for our approach is shown in Algorithm 13.

Algorithm 3: Algorithm for one iteration of the query selection strategy.

Data: Unlabeled instances U , Labeled instances L , Clusters C generated by event-based clustering, Number of instances to label per iteration b_i ,
 Trained Model for iteration M , Average size of all cluster in iteration ms ,
Result: Instances to label L

```

1 Use  $L$  to train classifier  $M$ ;
2 forall the clusters  $c$  in  $C$  do
3   forall the instances  $i$  in  $c$  do
4     Calculate information density  $DS(i)$  ;                               /* see Eq.23 */
5     Calculate entropy  $H(i)$  using  $M$  ;                               /* see Eq.25 */
6     Calculate density×entropy measure  $DSH(i)$  ;                       /* see Eq.26 */
7 forall the clusters  $c$  in  $C$  do
8   Calculate  $DSHC(c)$  ;                                               /* see Eq.27 */
9 Order clusters based on  $DSHC$ ;
10 while  $|L| \leq b_i$  do
11   forall the clusters  $c$  in  $C$  do
12      $n = \log_{ms}(|c|)$  ;                                           /* see Eq.28 */
13     Add  $n$  instances from  $c$  to  $L$ ;
```

In every iteration, the classifier trained on the currently labeled instances is applied to label all unlabeled instances. As a result, every instance is assigned a thematic dimension. Based on this, the event clustering is applied using the spatial, temporal, and thematic information resulting in a set of cluster C .

Next, for the query selection strategy, we calculate the information density DS per instance. For identifying informative instances, we use the instances for which the classifier trained on the currently labeled instances is most uncertain. As an uncertainty measure we use the entropy $H(x)$ [210], which is calculated for each instance x and each class $y \in Y = \{y_1, y_2, \dots, y_i\}$:

$$H(x) = - \sum_{y \in Y} P(y|x) \log P(y|x) \quad (25)$$

Based on the information density DS and the entropy $H(x)$, the *density × entropy* measure $DSH(x)$ [256] is calculated for each instance x :

$$DSH(x) = DS(x) \times H(x) \quad (26)$$

The informativeness and representativeness of each cluster is then calculated based on the average of DSH of each instance x in the cluster C :

$$DSHC(C) = \frac{\sum_{x \in C} DSH(x)}{N} \quad (27)$$

For selecting appropriate instances to query, the clusters are sorted by DSH of each cluster and the number of instances to draw per cluster is calculated using Formula 28.

$$n = \log_{ms} C \quad (28)$$

To determine how many (n) instances have to be selected per cluster, we calculate the average size of all cluster ms and the size of the current cluster C . We decided to use a logarithmic scale to avoid drawing too many instances from larger clusters: using a linear approach, large clusters would contribute more instances compared with small clusters. We assume that drawing only small numbers per cluster is sufficient as at some point additional instances will not yield any additional information. Furthermore, with our approach we draw a limited amount of instances per cluster to avoid choosing too similar instances for training as it would happen if $n = C/ms$ is used.

We select instances until the number of instances to label per iteration is reached. Based on the previous and the new instances, the model is retrained. The whole process is repeated until all iterations are finished.

In this section, we presented a novel active learning-based refinement strategy. We make use of temporal, spatial, and thematic information to perform an event-based clustering. Based on this approach, we are able to find a number of clusters for event classification that does not need to be specified beforehand compared with other clustering approaches. Our strategy allows for selecting initial training instances (Q1) as well as instances for refining the classifiers (Q2) taking informativeness and representativeness into account. As we show in the following section, our approach outperforms related work. Furthermore, it is less error prone to labeling quality.

9.4 Evaluation

In this section, we present the evaluation results regarding our proposed framework. We conducted four evaluations to show the advantage of our approach:

- In the first evaluation, we compared our approach with related approaches. This evaluation aimed at showing that event-based clustering outperforms other clustering-based active learning approaches.

-
- In a second evaluation, we evaluated the influence of noise on our approach as we showed that human-based classifications are not perfect (see Section 6.5.2).
 - Third, we evaluated the influence of different batch sizes and the size of the initial training set on our approach. This is important as a larger initial set leads to better results, whereas it is costly to obtain this set. In contrast, also frequently repeated querying for instances is expensive; thus, we were interested in the effects of both parameters, also with the varying number of annotators.
 - Fourth, we compared our approach with a plain supervised classification approach to get an idea of how fast our approach is able to yield comparable results.

9.4.1 Experimental Setup

Data Set and Classification: As there are no publicly available labeled data sets for event-related microposts, we needed to create our own high-quality ground truth data. For evaluating the event-based clustering, we used the data set SET_1_GT (see Section 6.5.2). The data set consists of 2,000 tweets with the following class distribution:

- 328 fire incidents, 309 crash incidents, 334 shooting incidents, 1,029 tweets related to no or other type of incident.

For our evaluation, we used 1,200 tweets from the data set for training and 800 tweets for testing. It is likely that the used data set violates the assumption of independently and identically distributed instances [102]. However, when it comes to active learning where labels for specific instances are requested by the learner, this assumption is violated inherently. Furthermore, all evaluated approaches have to deal with the same conditions, allowing us a fair comparison.

As a classifier, we used Weka’s implementation of John Platt’s sequential minimal optimization algorithm for training a support vector classifier [174]. The active learning algorithms select instances from the training set to query for labels. Based on these, a classifier was trained and then evaluated on the test set. Due to the complexity of determining the best parameter settings for each iteration and each approach, we follow related approaches (see [104] and [63]) and decided to compare all algorithms on fixed parameters ($c = 1$, polynomial kernel, $\epsilon = 10^{-12}$). Consequently, the SMO was used with standard settings even if tuning most likely would yield much better results. However, as we were interested in comparing different approaches, classification performance is not the primary goal. Also, as all approaches rely on the same parameter settings, we assume no advantage or disadvantage for one approach.

Furthermore, we did not conduct a feature selection for this classifier and chose word-3-grams and syntactic features as feature sets. Also, replacement strategies and semantic abstraction were applied. However, the missing feature selection is not supposed to influence the evaluation results as all approaches relied on the same feature set.

Metrics: To compare the different active learning approaches, we calculated the commonly used *deficiency* measure [19]. This measures the performance of an active learning algorithm throughout the whole learning session. We used the F1 Score based deficiency calculation [179] shown in Equation 29.

$$DEF(AL) = \frac{\sum_{t=1}^n (F1_n(REF) - F1_t(AL))}{\sum_{t=1}^n (F1_n(REF) - F1_t(REF))} \quad (29)$$

The deficiency is calculated using the achieved F1 score of the t -th iteration of a reference baseline algorithm (*REF*) and the compared active learning approach (*AL*). The result is normalized using the maximal F1 score and the learning curve of the reference algorithm *REF*. Thus, the measure is not negative, and values smaller than 1 indicate more efficient algorithms compared with the baseline strategy, whereas a value larger than 1 indicates a performance decrease in comparison with the reference approach.

Algorithms: In the evaluation, we applied different active learning algorithms. In order to evaluate the performance of our approach, we compared with two state-of-the-art clustering-based approaches, which also take representativeness and informativeness into account. Furthermore, we compared an entropy-based uncertainty approach that is commonly used in related works (see Section 9.2). The approaches reimplemented for comparison are the following:

- *Tang et al. [225]*: For initial sampling, a k -means clustering is used. For query selection, first, the most uncertain instances for each cluster are selected. Then information density is used to weight the examples. We set $k = 4$ because we have four event types in our classification problem.
- *Zhu et al. [256]*: For initial sampling, a k -means clustering is used (with $k = 4$). For selection criteria during the iterations, the *entropy* \times *density* measure is applied. In this case, no clustering is applied for the iterations.
- *Uncertainty*: Simple uncertainty sampling, which uses random instances, is applied for initial sampling. For query selection, the entropy-based uncertainty sampling [210] is used.

- *Our approach:* We apply the algorithms described in Section 9.3. Whenever clustering is conducted, we cluster all tweets within a spatial extent of 200 m and a temporal extent of 20 min.⁵⁷

Parameter Selection: Following the experimental settings of [102] and [104] we set the size of the initial training set as well as the size during the iterations to 50. No further tuning or parameterization was applied.

Error Rates: As one of our goals was to evaluate the effect of noise on our approach, we needed to quantify labeling quality. Thus, for our evaluations that take noise into account, we used the error rates we determined in the qualitative evaluation of human-based classifications in Section 6.5.2. Furthermore, it is important to take both error types (i.e., random and systematic errors) into account. Thus, we used 1,793 tweets for which the random error applies and 207 tweets for which the systematic error applies. Finally, we used the error rates determined in the corresponding study, which are 0.0331 for random error and 0.4247 for systematic error.

As humans are involved in all labeling tasks, we used these errors for the initial selection as well as for query selection. Based on the labels retrieved from all annotators for one instance, we applied majority voting to decide whether an instance gets a correct or a wrong label. In case the wrong label is chosen, we selected the second-best label determined in the prestudy. This way, we were able to cover the ambiguity aspect for tweets with systematic error. Take the following tweet for example:

CLEARED - firetrucks have left scene of reported carfire - NB 5 under convention center. That was quick

Nine of 14 participants assigned *fire incident*, and five assigned *car incident* as a label. As a correct label, the *fire incident* label was chosen. If the systematic error is applied, *car incident* is used as it is the second-best label. Thus, we avoided to assigning completely wrong labels to ambiguous tweets.

Furthermore, as the number of errors change with the number of annotators, we simulated different numbers. To get a better overview of the error rates, in Table 58, the error rates in relation to the number of annotators are shown. As displayed, the number of errors converges to zero with 200 or more annotators. Thus, we evaluated all approaches with 1, 5, 10, 20, 50, 100, and 200 annotators.

We conducted eight evaluations for each number of annotators and for the perfect case without error. Each evaluation for each algorithm and number of annotators was repeated 10 times as, for instance, the uncertainty approach is highly dependent on the selected instances. We report the averaged F1 score based on the repetitions.

⁵⁷ As a result, the 1,200 tweets of the training set are divided into 438 distinct event clusters.

Table 58.: Overview of the error rates with changing number of annotators (10 repetitions).

# of Annotators	1	5	10	20	50	100	200	300
# of Errors	62.9	24.1	14.6	10.6	6.5	2.4	0.4	0.0

9.4.2 Results

We conducted four evaluations: First, we compared with other state-of-the-art approaches using a perfect oracle and noisy annotators to answer Q1, Q2, and Q3. Second, we evaluated the influence of noise on our approach and, third, the influence of the batch size. Fourth, we compared our approach with plain supervised classification.

(1) Comparison to State-Of-The-Art Approaches:

The overall performance graphs and deficiencies are shown in Figure 50 and Figure 51. In Table 59, the deficiencies for all approaches are presented with the approach of Tang et al. as a baseline strategy.

Initial Selection Strategy: As can be seen in the figures, the performance after selecting the initial training set is always superior with our approach. Even with a small amount of initial instances, we already achieve a high F1 score. This shows that our approach seems to be more appropriate for selecting the initial data set.

Table 59.: Deficiencies $DEF(AL)$ of related strategies and our approach. The approach of Tang et al. is used as a baseline strategy.

Approach	No Noise	1 Annotator	5 Annotators	10 Annotators
Tang et al. [225]	1	1	1	1
Uncertainty	0.53	0.63	0.49	0.51
Zhu et al. [256]	0.90	0.93	0.83	0.87
Our Approach	0.44	0.53	0.39	0.44
	20 Annotators	50 Annotators	100 Annotators	200 Annotators
Tang et al. [225]	1	1	1	1
Uncertainty	0.53	0.55	0.53	0.53
Zhu et al. [256]	0.86	0.89	0.91	0.90
Our Approach	0.44	0.45	0.43	0.44

Query Selection Strategy: With respect to the performance of the iterations, our approach has a decreased deficiency compared with other clustering approaches (0.44 vs. 0.53). The approach of Zhu et al. outperforms the approach of Tang et al. in most cases and also with respect to the deficiency. We attribute this to the improved strategy for query selection. A surprising result is the performance of uncertainty sampling, which outperforms the other two clustering strategies. Apparently, only focusing on the informativeness seems to be a good strategy for our data set. In contrast, the number of distinct event types is used as the number of clusters might not be the most efficient approach. Also, uncertainty sampling performs well as the data set is rather small; thus, the effect of outliers might be low.

The graphs also show that our approach has a steep learning curve as, for instance, only a sixth of all instances are needed to achieve an F1 score of about 84%. This is especially important when it comes to labeling costs as only a limited amount of data would need to be labeled. One can see that our approach has a drop at 500 instances. This is most likely because at this point, the number of clusters is decreasing; thus, selecting appropriate instances is more difficult.

Robustness: In Table 59, the deficiencies for the different number of annotators are shown. With one annotator (i.e., the very error-prone case), the deficiencies are higher, thus worse compared with the case without errors. However, our event-based clustering still outperforms related strategies. The deficiencies decrease with an increasing number of annotators. In Figure 50 b) and 50 c), the learning curves for the very error-prone cases with one respective five annotators are shown. As can be seen in the learning curve of the approach of Tang et al., the influence of noise is notable in the big drop with 500 instances. Also the approach of Zhu et al. has a much lower initial F1 score compared with all other approaches, which is an indicator for an inappropriate initial selection strategy. The results indicate that even with noisy labels, our approach outperforms state-of-the-art approaches.

In all other cases, the learning curves are quite similar, which is a result of the decreased number of wrong labeled instances (see Table 51). The performance of all approaches increases with a lower number of errors. Also, the approaches of Tang et al. and Zhu et al. always have a low initial F1 score.

In summary, the approach of Zhu et al. always outperforms the approach of Tang et al., which is because of the improved selection strategy for query selection. Nevertheless, our approach outperforms all state-of-the-art approaches for all cases. Thus, we can conclude that event-based clustering that takes representative and informative instances into account is a promising strategy for active learning. Finally, we showed that our approach outperforms state-of-the-art for selecting an initial training set (Q1), for choosing appropriate instances for labeling in each iteration (Q2), as well as if labeling noise is taken into account (Q3).

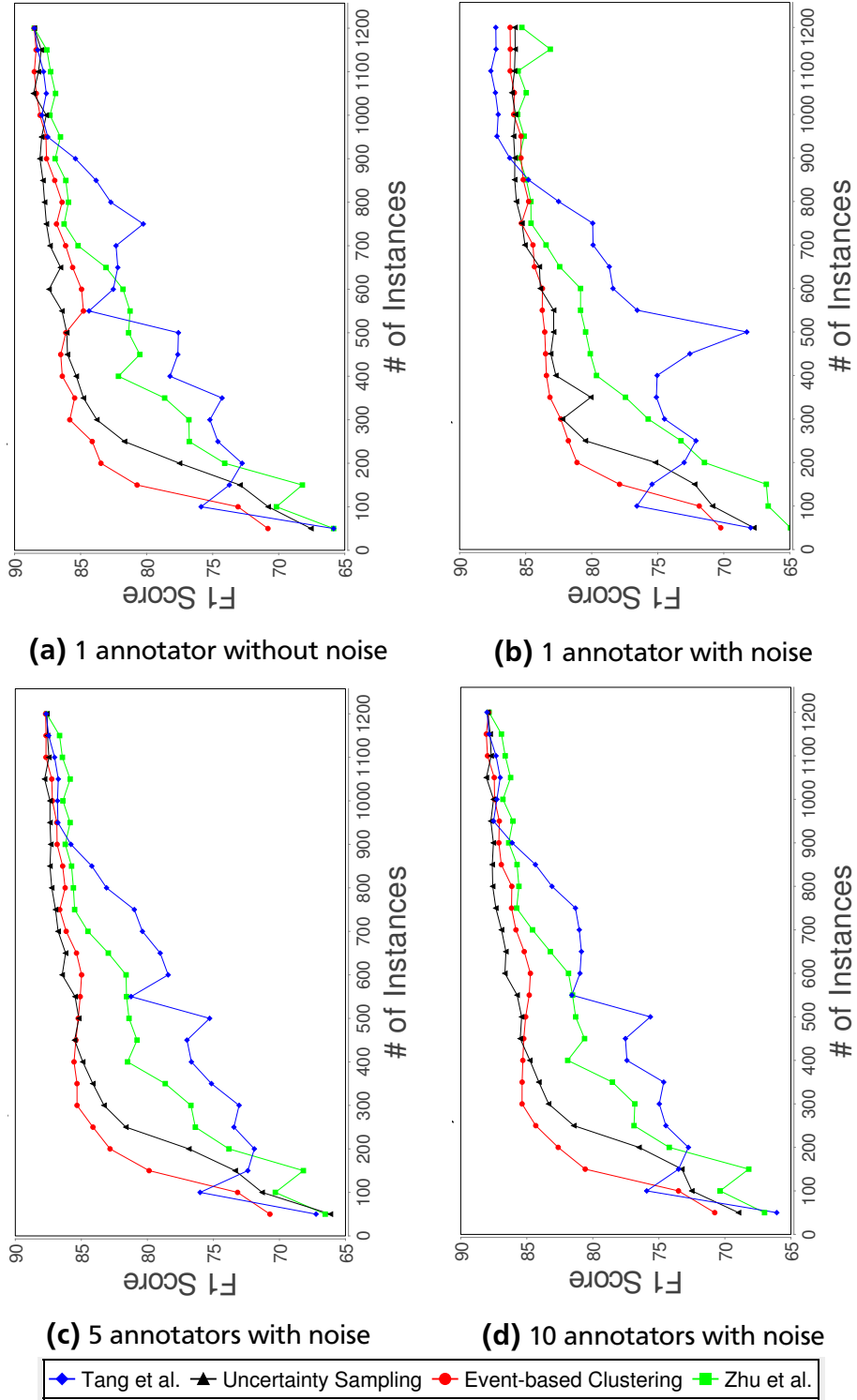


Figure 50.: Evaluation results of state-of-the-art selection strategies and our approach. The graphs for every combination of annotators with noise and without noise are shown.

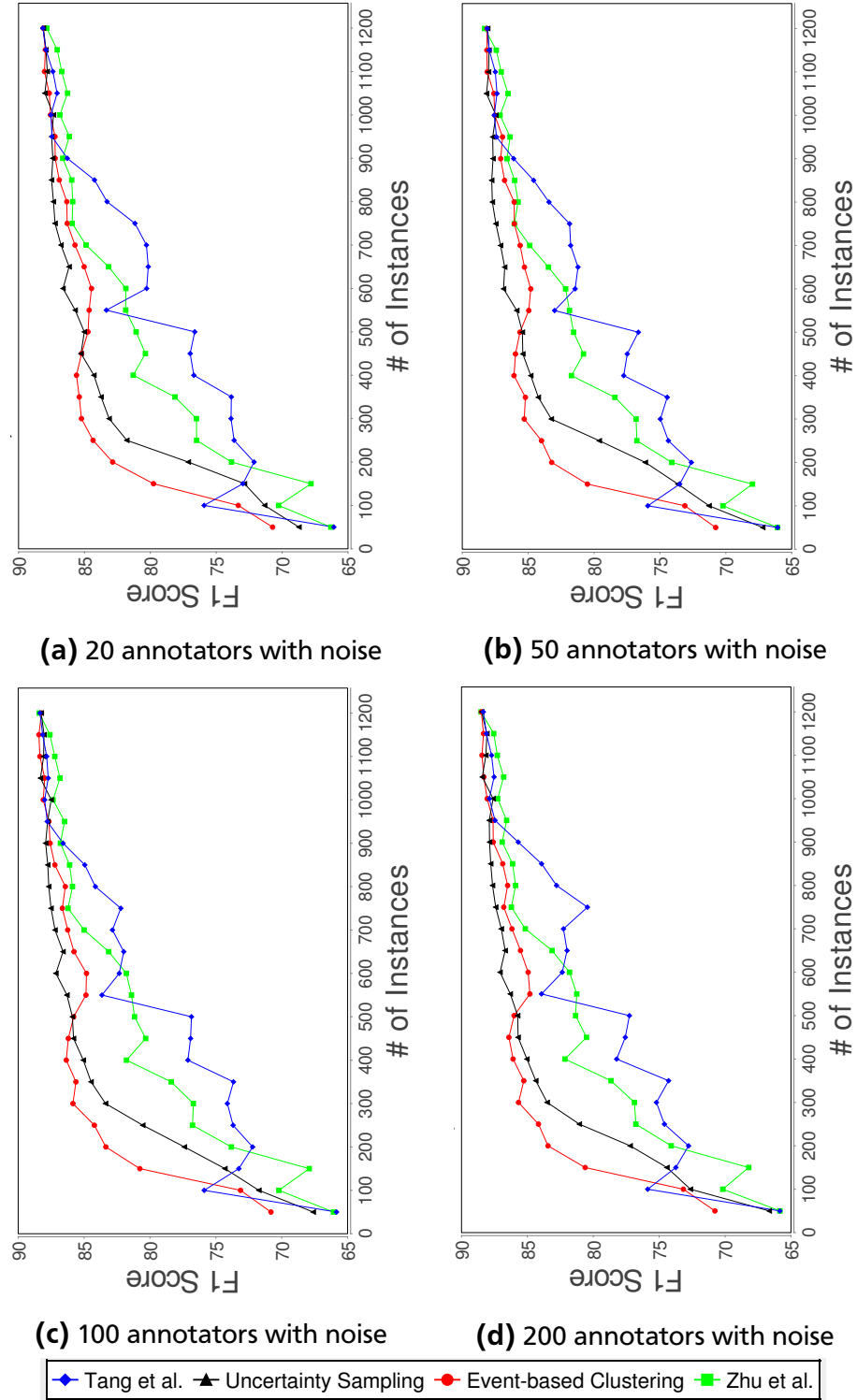


Figure 51.: Evaluation results of state-of-the-art selection strategies and our approach. The graphs for every combination of annotators with noise and without noise are shown.

(2) Influence of Noise:

With the first evaluation, we showed that our approach outperforms related work if noise is taken into account. As it is important to know how much annotators are needed to reduce noise to a minimum, we evaluated how noise influences our approach. In Figure 52, the graphs and Table 60 show the deficiency for our approach with a changing number of annotators. The approach without noise is used as a baseline strategy.

With an increasing number of annotators, noise is negligible. With only one annotator, the deficiency is worse by 57% and with 5 annotators still worse by 26%. With 10 to 50 annotators, the deficiency is worse by 10% to 12%. The graph also shows that for more than 10 annotators, an F1 score of 85% is reached comparably fast. With a maximum of five annotators, this level is only reached at the end of the simulation. For one annotator, this maximum is never achieved. These results indicate that a minimum number of annotators are needed for achieving good results by crowdsourced labeling tasks. In our experiments, 10 annotators seem to be sufficient, while in other domains with different error rates, there might be a need for much more annotators.

Table 60.: Deficiency of our approach with the number of annotators (number of errors in parentheses). No noise is used as a baseline strategy.

1	5	10	20	50	100	200
1.57 (63)	1.26 (23)	1.12 (14)	1.12 (10)	1.10 (7)	1.01 (2)	1.01 (0)

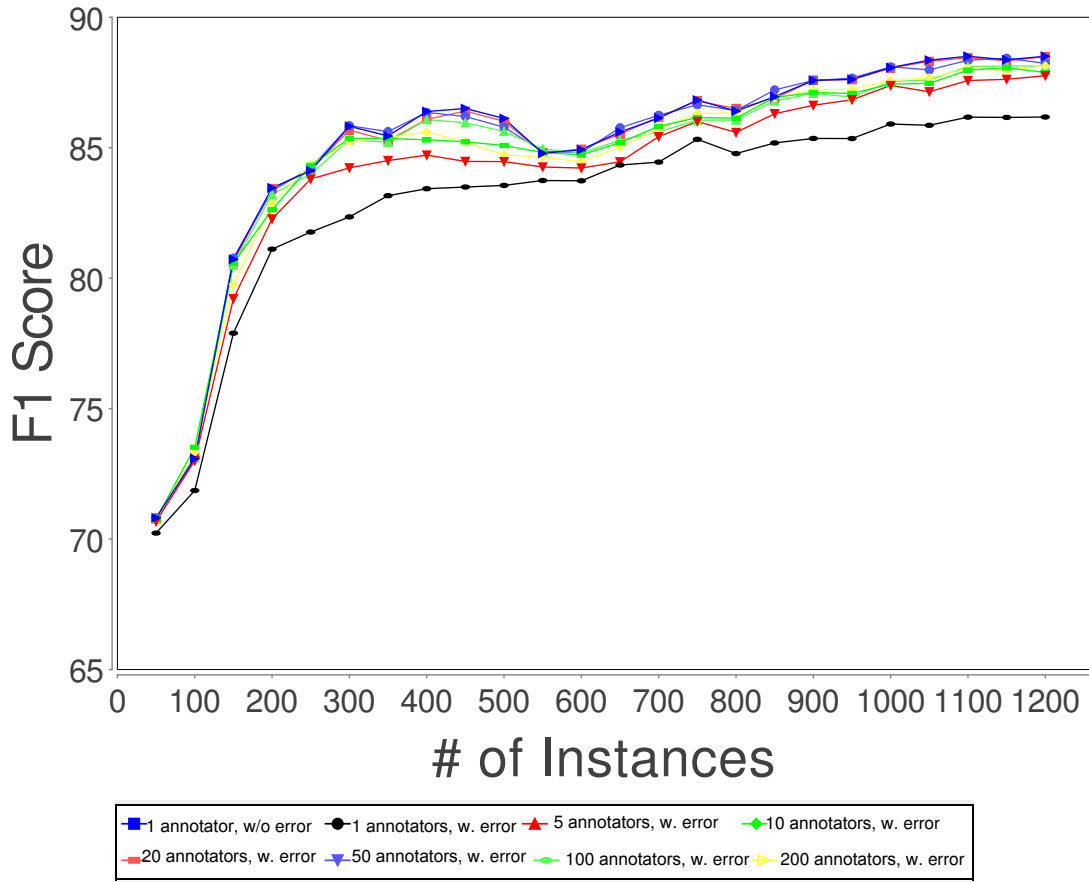


Figure 52.: Influence of noise on our approach. The graphs for every combination of annotators with and without noise are shown.

(3) Influence of Batch Size on Our Approach:

In Figure 53, the results of the evaluation of different batch sizes for our approach are shown. Not surprisingly, an increased number of selected instances per iteration leads to a faster improvement. As is shown in Figures 53 c) to e), a higher number of initial instances lead to much better initial performance. However, a high F1 score of 85% is reached in all approaches with approximately 500 labeled instances.

For the case without noise (200 annotators), the highest F1 score of about 88% is reached in all approaches with approximately 500 labeled instances. However, a high F1 score of 85% is achieved faster with 100 initial instances and 10 or 50 iteration instances respectively. This is because several iterations already took place; thus, more valuable instances have been chosen compared with approaches that start with 250 initial instances.

Furthermore, all figures underline previous results that a higher number of annotators (thus, lower amount of noise) are valuable for getting higher F1 scores for

all iterations as well as for the final F1 score. Starting with a higher number of initially labeled instances might be valuable as less iterations are needed. Nevertheless, taking more instances into account per iteration does not directly affect the performance of the classification. In this case, the appropriate strategy of choice is highly dependent on costs for setting up a labeling campaign. For example, if frequently repeated querying for instances is expensive, querying for a larger number of labels in fewer iterations might be beneficial.

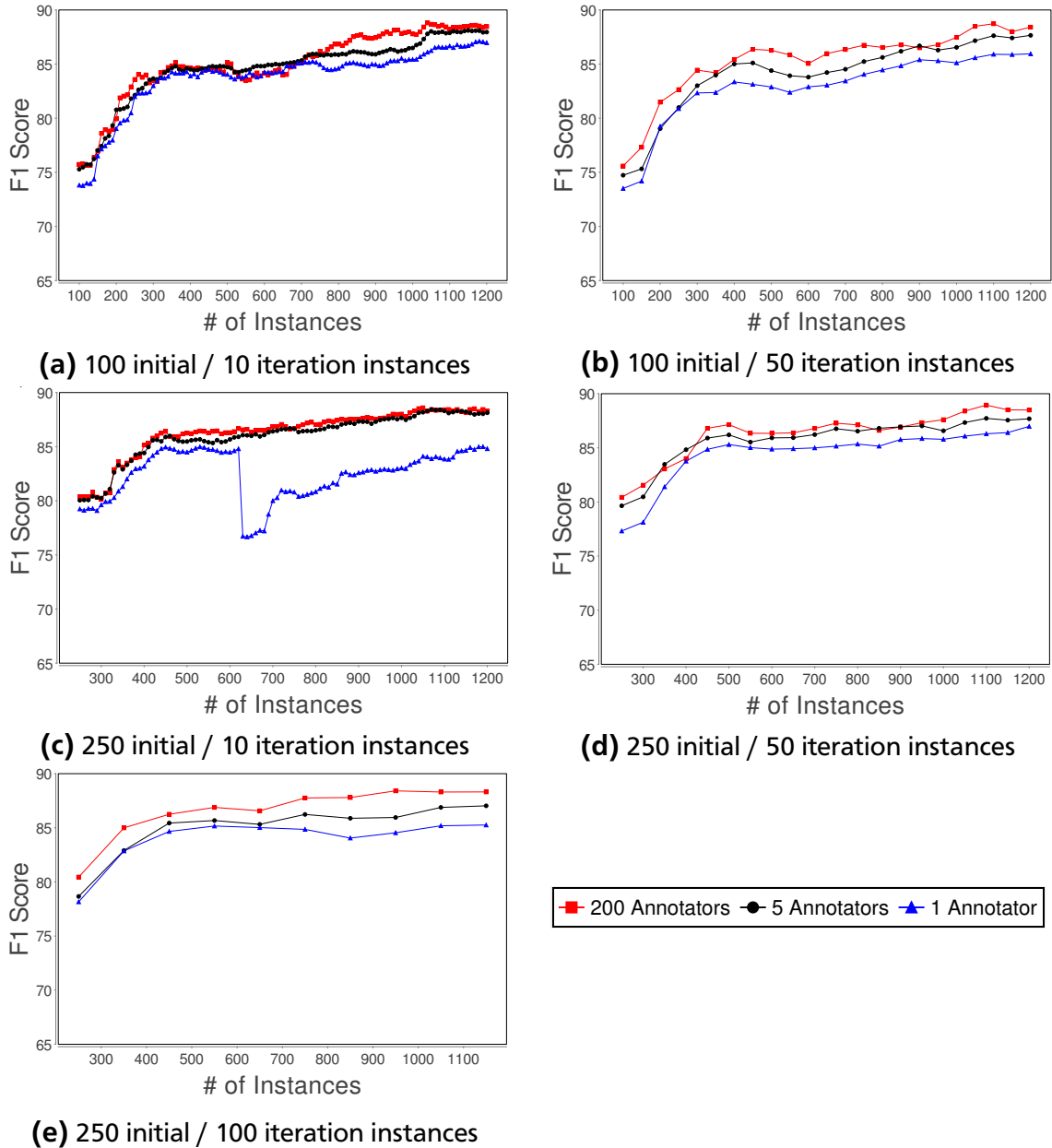


Figure 53.: The graphs for one, five, and 200 annotators are shown with different combinations of batch sizes.

(4) Comparison to Supervised Approach:

In the fourth evaluation, we compared our approach with a plain supervised approach, which takes all 1,200 instances into account. The supervised approach used the same feature set and classifier but used all instances at once for training. The approach was tested on the test set. Also, no noise was applied; thus, we assumed perfect labels. The classification results for the approach are shown in Table 61. With this approach taking all available data into account and not relying on noisy labels, an F1 score of 88.5% on the test set is achieved.

With our event-based clustering, we are able to achieve a comparable F1 score with the first iterations (about 300 instances). The maximum F1 score is reached after 19 iterations with 950 instances. Around 80% labeled instances are necessary to get the same F1 score as in the supervised approach, which shows that active learning can help to decrease labeling costs tremendously.

Table 61.: Classification results for all features of plain supervised approach that uses all instances (training on 1,200 instances, test on 800 instances).

Accuracy	Precision	Recall	F1
88.6%	88.6%	88.6%	88.5%

With event-based clustering, we are able to achieve a comparable F1 score of 85% after six iterations (300 instances, 5 annotators). The maximum F1 score with 88.4% is reached after 19 iterations with 950 instances. Around 80% labeled instances are necessary to get the same F1 score as in the supervised approach, which shows that active learning can help to decrease labeling costs tremendously. This underlines our initial hypothesis that active learning can help to significantly decrease labeling costs.

9.4.3 Summary

We conducted four evaluations and showed that our event-based clustering approach outperforms other clustering-based approaches. Furthermore, we showed that compared with a plain supervised learning approach, we need only one-third of the training data labeled to achieve comparable results.

9.5 Conclusion

In this chapter, we dealt with the challenge that social media platforms are dynamic environments; thus, classification of new data is necessary so that steps in the framework can be refined according to the changing conditions. Furthermore, this chapter addressed the challenge of high labeling costs and timely information retrieval on

the large amount of social media data, which is important in the emergency management domain. To solve this, we introduced an active learning approach based on event-based clustering. We demonstrated how information provided in event-related data such as tweets can be leveraged for the active learning process. For this process, we introduced a novel selection strategy based on temporal, spatial, and thematic information present in tweets. Our event-based clustering that identifies representative and informative instances outperforms state-of-the-art clustering approaches for active learning even with noisy labels.

- We presented two algorithms on the basis of event-based clustering for identifying (1) the initial training set and (2) the training sets for each iteration in active learning. In contrast to previous approaches, our algorithms leverage the temporal and spatial dimensions of event-related social media data to allow a more fine-grained clustering. Due to more natural and smaller clusters, our selection strategies are able to identify the most valuable instances for labeling.
- We validated the effectiveness of our approach on a data set of incident-related tweets compared with state-of-the-art approaches. We showed that our event-based clustering outperforms other clustering-based approaches for initial and for query selection.
- We tackled the commonly used assumption of perfect oracles in active learning environments and validated the influence of noise on our approach using quantified error rates. We showed that our approach is less prone to labeling errors compared with other approaches. For proofing this, we used error rates of crowdsourced annotators for evaluating the influence of noise.
- Our evaluation also showed that active learning is indeed valuable for reducing labeling costs of social media data. The evaluation of our approach showed that labeling costs can be reduced as fewer training data need to be labeled to achieve similar F1 scores compared with a supervised learner that has access to all training data.

For future work, several extensions might be provided. The event-based clustering framework could be used to allow the weighting of single features as these might highly be related to events. For instance, the n-gram "car_crash" might be an important feature for the class "crash". Allowing to weight features could lead to better results compared with an instance-based approach as features directly apply for multiple instances.

As outlined in Chapter 7.6, we also want to encode our data as a multi-label problem as incident-related tweets are ambiguous and can have more than one label at a time. For this, the effectiveness of our active learning approach needs to be evaluated. Furthermore, as we only demonstrated our approach on incident-related data, we also want to show the applicability on other types of events.

10 Conclusions

The goal of this dissertation was to answer the question *How can user-generated content be made a usable and valuable source of information for situational awareness of decision makers?* For this, we developed a framework that consists of the necessary steps to help answer this question. The framework is able to identify incidents based on user-generated content and to cluster incident-related information. This way, previously not usable user-generated content becomes a valuable source for emergency management.

In the following, the chapters of this dissertation are summarized, and the respective contributions are outlined. Furthermore, we point out directions of future research.

10.1 Summary

In the following, we summarize the main outcomes and contributions of this dissertation.

Part I - A Framework for Detecting and Clustering Incident Information in User-generated Content

In Chapter 2, we first introduced three essential requirements of a system that is able to identify incident-related information in user-generated content. Knowledge about the spatial, the temporal, and the thematic dimensions of an information item are inevitable for identifying incidents. Based on these requirements, we developed a framework for processing a large amount of user-generated content.

As a first step of the framework, user-generated content is collected. In Chapter 3, we described how an initial information base is created, which can further be processed in the subsequent steps of the framework. For this, we presented background on user-generated content and Twitter as our major source of incident-related information. Furthermore, we gave an overview of the data collection setup.

Part II - Automatic Preprocessing of User-Generated Content

To address the challenge that social media data is unstructured, we introduced several preprocessing steps in Chapter 4. These steps provided preliminaries needed for preparing text so it can be used for the subsequent steps of the framework. In the same chapter, we showed how named entities and temporal expressions are identified and extracted in order to use them as additional information when creating a machine learning model that is generalizable for data that stem from a different city.

For this, we presented a set of adaptations applied to standard techniques that allow us to extract named entities and temporal expressions from unstructured text. We also presented how we make use of the temporal expressions to infer the point in time when an incident occurred.

To satisfy the second requirement, we dealt with the problem of how to infer spatial information from user-generated content in Chapter 5. In this chapter, we contributed a novel approach for the geolocalization of tweets that is capable of inferring the home location of a Twitter user, the point of origin where a tweet was sent, as well as of inferring the location focus of a tweet message. We validated the accuracy of our approach and showed that the approach is able to locate 92% of all tweets with a median accuracy of below 30 km. Furthermore, it predicts the user's residence with a median accuracy of below 5.1 km. Finally, the same approach is able to estimate the focus of incident-related tweets within a median accuracy of below 250 m. This allows us to predict the spatial dimension of an incident with high precision.

Part III - Incident Detection and Clustering of Incident-Related Information

The goal of this part was to finally detect incidents based on user-generated content and to aggregate information related to the same incident. For this, we first contributed a general approach for applying crowdsourcing to classify and aggregate user-generated content according to the information need of the command staff in emergency management (see Chapter 6). With this approach, we are able to *manually* differentiate incident-related information from information not related to an incident. In this step, we also presented human-centered sensing as a means for collecting additional information about an incident. Our evaluation showed that this approach is indeed valuable for the command staff.

To address the problem that crowdsourcing is limited when it comes to the timely filtering of a large amount of information, we presented an approach for automatically detecting incident-related information in user-generated content (see Chapter 7). To develop a highly precise approach, we conducted an extensive evaluation of feature groups to determine an optimal set for this task. The performance evaluation showed that we are able to classify the incident type with an F-measure of more than 90%.

In the same chapter, we also dealt with the dynamism and regional variation of user-generated content. We contributed the novel concept of semantic abstraction, which allows creating features that are not city-specific and support training a generalized model. We evaluated semantic abstraction on data sets from five different cities and showed that it is indeed valuable for improving F-measure.

Based on the temporal, spatial, and thematic dimensions of each information item, we presented a clustering approach that is able to detect incidents in a large amount of social media data (see Chapter 8). The approach clusters all information related

to the same incident and is able to deal with different organizational incident type vocabularies. To underline the importance of our approach for emergency management, we presented the evaluation of the value of situational information shared in tweets, which were posted in two North American cities. We showed that a variety of individuals share information about small-scale incidents. Furthermore, we showed that important situational information about affected objects, injured persons, and the location of an incident is shared.

Closing the chapter, we evaluated the approach and showed that we are able to detect more than 50% of real-world incidents published in an official emergency management system. Furthermore, 32.14% of the detected incidents are within a 500 m radius and within a 10 min time interval of the real-world incident, allowing precise spatial and temporal localization. We showed that these results are more than five times better compared with related approaches. Also, more than 77% of the incident clusters created with our approach are indeed related to incidents.

As the machine learning approaches need to be adapted to changing conditions such as different incident types or different information sources, we presented an approach for refining the framework according to different information needs (see Chapter 9). For this, we introduced a novel event-based clustering approach that makes use of spatial, temporal, and thematic information. We validated the effectiveness of our approach on a data set of incident-related tweets compared with state-of-the-art approaches and showed that our approach outperforms related work. Also, the approach helps to reduce the amount of information that needs to be processed manually, thus reducing labeling costs tremendously in contrast to using a plain supervised learning approach.

The contributions of each individual chapter enable us finally to identify incidents based on a large amount of user-generated content. Furthermore, incident-related information is aggregated and now easily exploitable for decision making in emergency management.

10.2 Future Research Directions

As outlined in the individual chapters, there are many remaining opportunities for future work.

Our approach for the geolocalization of tweets presented in Chapter 5 currently does not use information provided in a user's social network for location inference. As users are highly interconnected, important location information can be derived using friendship relationships. Most importantly, calculating an overall confidence score for location estimation is another important aspect as this might help to avoid false predictions.

The human-based classification and aggregation step presented in Chapter 6 might be supported with additional prefiltering. For instance, named entities detected in the incident reports at hand are helpful for estimating the relationship between a report and a question. Thus, completely irrelevant information might be filtered out before it is classified by the crowd. Furthermore, additional and important background information might be provided using the links to external knowledge bases such as DBpedia.

Although the current approach for classifying incident-related tweets is rather precise, it can be adapted to handle more and different incident types. Furthermore, our approach of semantic abstraction could further be extended with other external knowledge bases. Also in this case, a comprehensive feature selection regarding the Linked Open Data features is needed to leverage the full potential of this information base. As outlined before, also applying multi-label classification is helpful as we already showed in [203]. This approach helps to assign multiple labels at once especially for the multi-annotation problem, thus losing less situational information. Furthermore, taking costs into account might help to avoid missing important information.

Several limitations of the rule-based clustering presented in Chapter 8 can be addressed in future work. First, an optimal rule set needs to be determined as we currently rely on simple rules. For this, interviews with emergency staff or automatic approaches for determining an optimal rule set could be conducted. Second, evaluating the vocabulary matching is open for future work.

Finally, the refinement step shown in Chapter 9 can also be extended. The event-based clustering approach could be used to allow weighting of single features as these might highly be related to events. Allowing to weight features could lead to better results compared with an instance-based approach as features directly apply for multiple instances.

In general, the whole framework can easily be refined to support different types of user-generated content from different sources. It might directly be applied to other textual content as it is shared on Facebook. Furthermore, the approach for the geolocalization of tweets and the machine-based classification can further be enhanced with more training data to achieve better results. Also, important information such as the number of injured people or affected buildings already present in the aggregated incident cluster could be detected automatically.

Furthermore, the pipeline needs to be evaluated in a long-term run and with respect to real-time applicability. For emergency management, it is necessary to retrieve important information as fast as possible; thus, our framework needs further tuning toward this aspect. The main bottleneck is likely the geolocalization part as many different and external APIs are used for precise estimations. In comparison, applying the machine learning model and creating incident clusters are not as time intensive.

Also, investigating privacy issues is open for future research. In particular, our approach for geolocalization makes use of sensitive metadata about a user. Protecting privacy and concurrently gathering important incident-related information is a remaining challenge for data mining on social media.

As a result of this dissertation, we presented several contributions that help to identify incidents based on user-generated content and to cluster incident-related information. This way, previously not usable user-generated content becomes a valuable source for decision making in emergency management.



A Appendix

A.1 Evaluation of Semantic Abstraction in Addition to the Best Feature Groups

In Table 62, Table 63, and Table 64 the results for applying the individual approaches of semantic abstraction compared to not using them are shown. The results are presented for the 4-CLASSES data set.

Table 62.: Comparison of using semantic abstraction compared to a baseline comprising n-gram features after Slang and URL replacement, TF-IDF scores, and syntactic features on 2-CLASSES data set with word-2-grams and SVM as classifier.

Features	c	Accuracy	F-measure	w/o Sem.Abstr.
Baseline (binary)		64.67%	50.79%	
+SLANG +URL	0.5	90.05%	89.98%	
+TF-IDF +SYNT				
+ALL	0.03125	90.47%	90.41%	0.44%
+LOC	0.125	90.66%	90.60%	0.62%
+TIME	0.125	90.17%	90.07%	0.10%
+LOD	0.03125	89.62%	89.53%	-0.44%
+LOC+TIME	0.125	90.78%	90.71%	0.73%
+LOC + LOD	0.03125	90.38%	90.32%	0.34%
+TIME + LOD	0.03125	89.59%	89.50%	-0.47%
+TYPES	0.03125	89.77%	89.67%	-0.31%
+CATEGORIES	0.03125	89.65%	89.54%	-0.43%

Table 63.: Comparison of using semantic abstraction compared to a baseline comprising n-gram features after Slang and URL replacement, TF-IDF scores, and syntactic features on 2-CLASSES data set with word-3-grams and SVM as classifier.

Features	c	Accuracy	F-measure	w/o Sem.Abstr.
Baseline (binary)		64.67%	50.79%	
+SLANG +URL	0.5	89.93%	89.81%	
+TF-IDF +SYNT				
+ALL	0.03125	90.08%	90.00%	0.19%
+LOC	0.03125	90.60%	90.49%	0.68%
+TIME	0.5	90.14%	90.03%	0.22%
+LOD	0.03125	89.65%	89.55%	-0.26%
+LOC+TIME	0.03125	90.66%	90.55%	0.74%
+LOC + LOD	0.03125	90.32%	90.24%	0.43%
+TIME + LOD	0.03125	89.50%	89.40%	-0.41%
+TYPES	0.03125	89.77%	89.65%	-0.16%
+CATEGORIES	0.03125	89.65%	89.53%	-0.28%

Table 64.: Comparison of using semantic abstraction compared to a baseline comprising n-gram features after Slang and URL replacement, TF-IDF scores, and syntactic features on 2-CLASSES data set with char-5-grams and NB as classifier.

Features	c	Accuracy	F-measure	w/o Sem.Abstr.
Baseline		64.67%	50.79%	
+SLANG +URL		88.86%	88.83%	
+TF-IDF +SYNT				
+ALL		88.15%	88.08%	-0.75%
+LOC		88.05%	87.98%	-0.84%
+TIME		88.71%	88.68%	-0.15%
+LOD		88.90%	88.80%	-0.02%
+LOC+TIME		87.95%	87.87%	-0.96%
+LOC + LOD		88.00%	87.93%	-0.90%
+TIME + LOD		88.90%	88.80%	-0.03%
+TYPES		89.30%	89.19%	0.37%
+CATEGORIES		89.15%	89.03%	0.20%

In Table 65, Table 66, and Table 67 the results for applying the individual approaches of semantic abstraction compared to not using them are shown. The results are presented for the 4-CLASSES data set.

Table 65.: Comparison of using semantic abstraction compared to a baseline comprising n-gram features after Slang and URL replacement, TF-IDF scores, and syntactic features on 4-CLASSES data set with word-2-grams and SVM as classifier.

Features	c	Accuracy	F-measure	w/o Sem.Abstr.
Baseline (binary)	0	51.45%	34.96%	
+SLANG +URL	8	92.05%	92.02%	
+TF-IDF +SYNT				
+ALL	0.03125	90.15%	90.10%	-1.92%
+LOC	0.03125	91.90%	91.86%	-0.16%
+TIME	0.125	92.15%	92.11%	0.09%
+LOD	0.03125	90.40%	90.35%	-1.67%
+LOC+TIME	0.03125	91.75%	91.71%	-0.30%
+LOC + LOD	0.03125	90.30%	90.24%	-1.78%
+TIME + LOD	0.125	90.15%	90.10%	-1.92%
+TYPES	512	90.80%	90.76%	-1.26%
+CATEGORIES	0.125	90.95%	90.91%	-1.11%

Table 66.: Comparison of using semantic abstraction compared to a baseline comprising n-gram features after Slang and URL replacement, TF-IDF scores, and syntactic features on 4-CLASSES data set with word-3-grams and SVM as classifier.

Features	c	Accuracy	F-measure	w/o Sem.Abstr.
Baseline (binary)	0	51.45%	34.96%	
+SLANG +URL	512	92.00%	91.95%	
+TF-IDF +SYNT				
+ALL	0.125	90.10%	90.05%	-1.90%
+LOC	0.125	91.90%	91.87%	-0.08%
+TIME	8	92.05%	92.00%	0.05%
+LOD	0.03125	90.35%	90.29%	-1.66%
+LOC+TIME	2048	91.75%	91.72%	-0.23%
+LOC + LOD	128	90.25%	90.22%	-1.73%
+TIME + LOD	0.03125	90.10%	90.05%	-1.91%
+TYPES	0.03125	90.70%	90.64%	-1.32%
+CATEGORIES	0.03125	91.00%	90.94%	-1.01%

Table 67.: Comparison of using semantic abstraction compared to a baseline comprising n-gram features after Slang and URL replacement, TF-IDF scores, and syntactic features on 4-CLASSES data set with char-5-grams and NB as classifier.

Features	c	Accuracy	F-measure	w/o Sem.Abstr.
Baseline		51.45%	34.96%	
+SLANG +URL		88.85%	88.74%	
+TF-IDF +SYNT				
+ALL		88.15%	88.08%	-0.66%
+LOC		88.05%	87.98%	-0.76%
+TIME		88.45%	88.32%	-0.43%
+LOD		88.90%	88.80%	0.06%
+LOC+TIME		87.95%	87.87%	-0.87%
+LOC + LOD		88.00%	87.93%	-0.81%
+TIME + LOD		88.90%	88.80%	0.06%
+TYPES		89.30%	89.19%	0.45%
+CATEGORIES		89.15%	89.03%	0.29%

A.2 Evaluation Results of Machine-Based Aggregation of User-Generated Content

In Table 68 and Table 69 the evaluation results of machine-based aggregation using the leader-following and the DBScan clustering algorithms are shown. In the tables, the recall of detecting real-world incidents within a 500 m radius is shown. Furthermore, the accuracy (i.e., the number of incident-related clusters compared to all created clusters) is shown.

Table 68.: Evaluation results of machine-based aggregation using leader-follower clustering.

Threshold	# Clusters	Ø Size	Incident Related	Rec.(< 500m)	Acc.
0.25	136	1	3	3.53%	2.21%
0.5	135	3	3	3.53%	2.22%
0.75	133	4	4	4.71%	3.01%
0.8	132	5	5	5.88%	3.79%
0.9	127	8	4	4.71%	3.15%
1.0	111	7	4	4.71%	3.60%
1.1	78	2	3	3.53%	3.85%
1.2	63	2	1	1.18%	1.59%
1.25	53	2	0	0.00%	0.00%
1.5	34	3	2	2.35%	5.88%
2.0	21	373	1	1.18%	4.76%

Table 69.: Evaluation results of machine-based aggregation using DBScan clustering.

MinPoints	Epsilon	# Clusters	Ø Size	Incident Related	Rec.(< 500m)	Acc.
1	0.005	4	934	2	2.35%	50.00%
2	0.005	2	1988	1	1.18%	50.00%
5	0.005	2	1988	1	1.18%	50.00%
1	0.01	4	934	2	2.35%	50.00%
2	0.01	2	1988	1	1.18%	50.00%
5	0.01	2	1988	1	1.18%	50.00%
1	0.015	4	934	2	2.35%	50.00%
2	0.015	2	1988	1	1.18%	50.00%
5	0.015	2	1988	1	1.18%	50.00%
1	0.02	4	934	2	2.35%	50.00%
2	0.02	2	1988	1	1.18%	50.00%
5	0.02	2	1988	1	1.18%	50.00%
1	0.025	4	934	2	2.35%	50.00%
2	0.025	2	1988	1	1.18%	50.00%
5	0.025	2	1988	1	1.18%	50.00%
1	0.05	4	934	2	2.35%	50.00%
2	0.05	2	1988	1	1.18%	50.00%
5	0.05	2	2015	1	1.18%	50.00%
1	0.25	4	926	2	2.35%	50.00%
2	0.25	2	1981	1	1.18%	50.00%
5	0.25	2	2107	1	1.18%	50.00%
1	0.4	8	478	3	3.53%	37.50%
2	0.4	4	1038	1	1.18%	25.00%
3	0.4	2	2157	1	1.18%	50.00%
4	0.4	2	2197	1	1.18%	50.00%
5	0.4	2	2227	1	1.18%	50.00%
1	0.5	11	647	3	3.53%	27.27%
2	0.5	4	1041	1	1.18%	25.00%
3	0.5	3	1450	1	1.18%	33.33%
4	0.5	2	2217	1	1.18%	50.00%
5	0.5	2	2247	1	1.18%	50.00%
1	0.6	7	543	2	2.35%	28.57%
2	0.6	3	1396	2	2.35%	66.67%
3	0.6	3	1458	2	2.35%	66.67%
4	0.6	3	1489	2	2.35%	66.67%
5	0.6	3	1508	2	2.35%	66.67%
1	0.75	5	536	1	1.18%	20.00%
2	0.75	1	2916	1	1.18%	100.00%
5	0.75	1	3049	1	1.18%	100.00%
1	1	0	0	0	0.00%	0.00%
2	1	0	0	0	0.00%	0.00%
5	1	0	0	0	0.00%	0.00%

Bibliography

- [1] Abel, F., Hauff, C., Houben, G.-J., Stronkman, R., and Tao, K. Twitcident: Fighting Fire with Information from Social Web Stream. In *Proceedings of the 21st World Wide Web Conference, WWW 2012*, pages 305–308. ACM, 2012.
- [2] Abel, F., Hauff, C., and Stronkman, R. Semantics + Filtering + Search = Twitcident Exploring Information in Social Web Streams Categories and Subject Descriptors. In *Proceedings of the 23rd ACM Conference on Hypertext and Social Media*, pages 285–294. ACM, 2012.
- [3] Abrol, S. and Khan, L. Tweethood: Agglomerative Clustering on Fuzzy k-Closest Friends with Variable Depth for Location Mining. In *Proceedings of the 2010 IEEE Second International Conference on Social Computing, SOCIALCOM '10*, pages 153–160. IEEE Computer Society, 2010.
- [4] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media, LSM '11*, pages 30–38. Association for Computational Linguistics, 2011.
- [5] Agarwal, P., Vaithyanathan, R., Sharma, S., and Shroff, G. Catching the long-tail: Extracting local news events from twitter. In *Proceedings of the Sixth International Conference on Weblogs and Social Media, ICWSM 2012*. AAAI Press, 2012.
- [6] Aggarwal, C. C. and Zhai, C. X. *Mining Text Data*. Springer-Verlag, 2012.
- [7] Ahmed, A., Hong, L., and Smola, A. J. Hierarchical geographical modeling of user locations from social media posts. In *Proceedings of the 22nd International Conference on World Wide Web, WWW'13*, pages 25–36. International World Wide Web Conferences Steering Committee, 2013.
- [8] Ahn, D., van Rantwijk, J., and de Rijke, M. A cascaded machine learning approach to interpreting temporal expressions. In *Proceedings of the 2007 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, HLT-NAACL*, pages 420–427. The Association for Computational Linguistics, 2007.
- [9] Ajmani, V. Los Angeles Fire Tweets. Online, 2008. URL <http://www.mibazaar.com/lafires.php>. Accessed: 01.12.2012.
- [10] Ajmani, V. Swine Flu Tweets. Online, 2009. URL <http://www.mibazaar.com/swineflu.html>. Accessed: 01.12.2012.

-
-
- [11] Ajmani, V. Iran Election protests. Online, 2009. URL <http://www.mibazaar.com/irantweets.html>. Accessed: 01.12.2012.
- [12] Albakour, M.-D., Macdonald, C., and Ounis, I. Identifying local events by using microblogs as social sensors. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval, OAIR '13*, pages 173–180. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 2013.
- [13] Allan, J., Carbonell, J., Doddington, G., Yamron, J., and Yang, Y. Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218, 1998.
- [14] Allan, J., Papka, R., and Lavrenko, V. On-line new event detection and tracking. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '98*, pages 37–45. ACM, 1998.
- [15] Amitay, E., Har'El, N., Sivan, R., and Soffer, A. Web-a-Where : Geotagging Web Content. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '04*, pages 273–280. ACM, 2004.
- [16] Asur, S. and Huberman, B. A. Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT '10*, pages 492–499. IEEE Computer Society, 2010.
- [17] Bailey, P., Thomas, P., Craswell, N., Vries, A. P. D., Soboroff, I., and Yilmaz, E. Relevance assessment: are judges exchangeable and does it matter. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '08*, pages 667–674. ACM, 2008.
- [18] Balcan, M.-F., Broder, A. Z., and Zhang, T. Margin based active learning. In *Proceedings of 20th Annual Conference on Learning Theory, COLT'07*, volume 4539 of *Lecture Notes in Computer Science*, pages 35–50. Springer-Verlag, 2007.
- [19] Baram, Y., El-Yaniv, R., and Luz, K. Online choice of active learning algorithms. *Journal of Machine Learning Research*, 5:255–291, 2004.
- [20] Beaumont, C. New york plane crash: Twitter breaks the news, again. Online, 2009. URL <http://www.telegraph.co.uk/technology/twitter/4269765/>



- New-York-plane-crash-Twitter-breaks-the-news-again.html. Accessed: 01.03.2014.
- [21] Becker, H. *Identification and Characterization of Events in Social Media*. PhD thesis, Columbia University, 2011.
- [22] Becker, H., Naaman, M., and Gravano, L. Beyond trending topics: Real-world event identification on twitter. Technical report, Columbia University, 2011.
- [23] Becker, H., Naaman, M., and Gravano, L. Beyond trending topics: Real-world event identification on twitter. In *Fifth International AAAI Conference on Weblogs and Social Media, ICWSM'11*. AAAI Press, 2011.
- [24] Berberich, K., Bedathur, S., Alonso, O., and Weikum, G. A language modeling approach for temporal information needs. In *Proceedings of the 32nd European Conference on Information Retrieval, ECIR 2010*, pages 13–25. Springer-Verlag, 2010.
- [25] Berlingerio, M., Calabrese, F., Lorenzo, G. D., Dong, X., Gkoufas, Y., and Mavroeidis, D. Safercity: A system for detecting and analyzing incidents from social media. In *Proceedings of 13th IEEE International Conference on Data Mining Workshops, ICDMW'13*, pages 1077–1080. IEEE Computer Society, 2013.
- [26] Berners-Lee, T. Linked data. Online, 2006. URL <http://www.w3.org/DesignIssues/LinkedData.html>. Accessed: 01.02.2014.
- [27] Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. DBpedia - A crystallization point for the Web of Data. *Web Semantics - Science Services and Agents on the World Wide Web*, 7(3):154–165, 2009.
- [28] Blanchard, W. Select emergency management-related terms and definitions. Online, 2006. URL <http://www.hSDL.org/?view&did=480235>. Accessed: 01.04.2014.
- [29] Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.
- [30] Blitzer, J., McDonald, R., and Pereira, F. Domain adaptation with structural correspondence learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP'06*, pages 120–128. Association for Computational Linguistics, 2006.
- [31] Boettcher, A. and Lee, D. Eventradar: A real-time local event detection scheme using twitter stream. In *Proceedings of the 2012 IEEE International*

Conference on Green Computing and Communications, GREENCOM '12, pages 358–367. IEEE Computer Society, 2012.

- [32] Bouilliot, F., Poncelet, P., and Roche, M. How and why exploit tweet's location information? In *Proceedings of 15th AGILE International Conference on Geographic Information Science, AGILE'12*, pages 3–7. Association of Geographic Information Laboratories for Europe, 2012.
- [33] Brants, T., Chen, F., and Farahat, A. A system for new event detection. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, SIGIR '03*, pages 330–337. ACM, 2003.
- [34] Brownstein, J. and Freifeld, C. Healthmap: the development of automated real-time internet surveillance for epidemic intelligence. *Eurosurveillance*, 12 (48):3322–3324, 2007.
- [35] Burke, J., Estrin, D., Hansen, M., Parker, A., Ramanathan, N., Reddy, S., and Srivastava, M. B. Participatory sensing. In *Proceedings of the Workshop on World-Sensor-Web: Mobile Device Centric Sensor Networks and Applications, WSW'06*, pages 117–134. ACM, 2006.
- [36] Campbell, A. T., Eisenman, S. B., Lane, N. D., Miluzzo, E., Peterson, R. A., Lu, H., Zheng, X., Musolesi, M., Fodor, K., and Ahn, G.-S. The rise of people-centric sensing. *IEEE Internet Computing*, 12(4):12–21, 2008.
- [37] Cano, A. E., Varga, A., Rowe, M., Ciravegna, F., and He, Y. Harnessing linked knowledge sources for topic classification in social media. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media, HT '13*, pages 41–50. ACM, 2013.
- [38] Carvalho, L. S., S. and Rossetti, R. Real-time sensing of traffic information in twitter messages. In *Proceedings of the 4th Workshop on Artificial Transportation Systems and Simulation ATSS, ITSC'10*, pages 19–22. IEEE Computer Society, 2010.
- [39] Cataldi, M., Di Caro, L., and Schifanella, C. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining, MDMKDD '10*, pages 1–10. ACM, 2010.
- [40] Centre for Monitoring Election Violence. Election day violence. Online, 2008. URL <http://cmev.wordpress.com/maps/>. Accessed: 01.04.2013.
- [41] Chae, J., Maciejewski, R., Bosch, H., Thom, D., Jang, Y., Ebert, D. S., and Ertl, T. Spatiotemporal social media analytics for abnormal event detection and



- examination using seasonal-trend decomposition. In *Proceedings of the 2012 IEEE Conference on Visual Analytics Science and Technology, VAST '12*, pages 143–152. IEEE Computer Society, 2012.
- [42] Chandra, S., Khan, L., and Muhaya, F. B. Estimating twitter user location using social interactions - a content based approach. In *2011 IEEE Third Intl Conference on Privacy Security Risk and Trust and 2011 IEEE Third Intl Conference on Social Computing, SocialCom'11*, pages 838–843. IEEE Computer Society, 2011.
- [43] Chang, H.-W., Lee, D., Eltaher, M., and Lee, J. @phillies tweeting from philly? predicting twitter user locations with spatial word usage. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining, ASONAM'12*, pages 111–118. IEEE Computer Society, 2012.
- [44] Chang, H.-C. A new perspective on twitter hashtag use: Diffusion of innovation theory. In *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem, ASIS&T '10*, volume 47, pages 1–4. American Society for Information Science, 2010.
- [45] Cheng, Z., Caverlee, J., and Lee, K. You are where you tweet : A content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management, CIKM '10*, pages 759–768. ACM, 2010.
- [46] Chinchor, N. A. Named entity task definition. In *Proceedings of the Seventh Message Understanding Conference, MUC-7*, 1998. URL http://acl.ldc.upenn.edu/muc7/ne_task.html.
- [47] Clausen, J. Live earthquake mashup. Online, 2012. URL <http://www.oe-files.de/gmaps/eqmashup.html>. Accessed: 01.12.2012.
- [48] Clodoveu, A. D. J., Pappa, G. L., de Oliveira, D. R. R., and de L. Arcanjo, F. Inferring the location of twitter messages based on user relationships. *Transactions in GIS*, 15(6):735–751, 2011.
- [49] Cohen, W. W. Fast effective rule induction. In *In Proceedings of the Twelfth International Conference on Machine Learning, ICML'95*, pages 115–123. Morgan Kaufmann Publishers Inc., 1995.
- [50] Cohn, D. A., Ghahramani, Z., and Jordan, M. I. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.
- [51] Collins, A. and Loftus, E. A spreading-activation theory of semantic processing. *Psychological review*, 82(6):407–428, 1975.

-
-
- [52] Conover, M., Gonçalves, B., Ratkiewicz, J., Flammini, A., and Menczer, F. Predicting the political alignment of twitter users. In *Proceedings of Third international conference on Social Computing, SocialCom'11*, pages 192–199. IEEE Computer Society, 2011.
- [53] Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Flammini, A., and Menczer, F. Political polarization on twitter. In *Proceedings of 5th International AAAI Conference on Weblogs and Social Media, ICWSM'11*. AAAI Press, 2011. URL <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2847>.
- [54] Cordeiro, M. Twitter event detection: Combining wavelet analysis and topic inference summarization. In *Doctoral Symposium on Informatics Engineering, DSIE*, 2012.
- [55] Cristianini, N. and Shawe-Taylor, J. *An Introduction to Support Vector Machines: And Other Kernel-based Learning Methods*. Cambridge University Press, 2000.
- [56] Dasgupta, S. and Hsu, D. Hierarchical sampling for active learning. In *Proceedings of the 25th International Conference on Machine Learning, ICML'08*, pages 208–215. ACM, 2008.
- [57] Daumé, H., III. Frustratingly easy domain adaptation. In *Proceedings of 45th Annual Meeting of the Association for Computational Linguistics, ACL'07*, pages 256–263. Association for Computational Linguistics, 2007.
- [58] Daumé, H., III and Marcu, D. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26(1):101–126, 2006. ISSN 1076-9757.
- [59] De Choudhury, M., Diakopoulos, N., and Naaman, M. Unfolding the event landscape on twitter: classification and exploration of user categories. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, CSCW'12*, pages 241–244. ACM, 2012.
- [60] Ding, J., Gravano, L., and Shivakumar, N. Computing geographical scopes of web resources. In *Proceedings of the 26th International Conference on Very Large Databases, VLDB '00*, pages 545–556. Morgan Kaufmann Publishers Inc., 2000.
- [61] Do, C. and Ng, A. Y. Transfer learning for text classification. In *Proceeding of the Advances in Neural Information Processing Systems, NIPS'05*, 2005. URL <http://dblp.uni-trier.de/db/conf/nips/nips2005.html#DoN05>.

-
- [62] Donmez, P. and Carbonell, J. G. Proactive learning: cost-sensitive active learning with multiple imperfect oracles. In *Proceedings of the 17th ACM conference on Information and knowledge management, CIKM '08*, pages 619–628. ACM, 2008.
- [63] Donmez, P., Carbonell, J. G., and Bennett, P. N. Dual strategy active learning. In *Proceedings of the 18th European Conference on Machine Learning, ECML'07*, pages 116–127. Springer-Verlag, 2007.
- [64] Donmez, P., Carbonell, J. G., and Schneider, J. Efficiently learning the accuracy of labeling sources for selective sampling. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'09*, pages 259–268. ACM, 2009.
- [65] Donmez, P., Carbonell, J. G., and Schneider, J. G. A probabilistic framework to learn from multiple annotators with time-varying accuracy. In *Proceedings of the SIAM International Conference on Data Mining, SDM'10*, pages 826–837. SIAM, 2010.
- [66] Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP '10*, pages 1277–1287. Association for Computational Linguistics, 2010.
- [67] emaps.com online maps of Australia, A. Bushfire incidents. Online, 2012. URL <http://www.aus-emaps.com/fires.php>. Accessed: 01.12.2012.
- [68] Endsley, M. R. Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors Society 32nd Annual Meeting*, volume 32 of *Aerospace Systems: Situation Awareness in Aircraft Systems*, pages 97–101. Human Factors and Ergonomics Society, 1988.
- [69] Ester, M., peter Kriegel, H., Sander, J., and Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96*, pages 226–231. AAAI Press, 1996.
- [70] Everitt, B. *The analysis of contingency tables*. Chapman and Hall, 2nd edition, 1977.
- [71] Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9: 1871–1874, 2008.
- [72] Federal Emergency Management Agency. FEMA 2014 Strategic Plan Overview. Online, 2011. URL http://www.fema.gov/about/2011_14_strategic_plan_overview.shtm. Accessed: 01.04.2013.

-
- [73] Feldman, R. and Sanger, J. *The Text Mining Handbook - Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, 2007.
- [74] Finkel, J. R., Grenager, T., and Manning, C. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL05*, pages 363–370. Association for Computational Linguistics, 2005.
- [75] Frakes, W. B. and Baeza-Yates, R. *Information Retrieval: Data Structures and Algorithms*. Prentice-Hall, Inc., 1992.
- [76] Freund, Y., Seung, H. S., Shamir, E., and Tishby, N. Selective sampling using the query by committee algorithm. *Machine Learning Journal*, 28(2-3):133–168, 1997.
- [77] Fritz, C., Kirschner, C., Reker, D., and Wisplinghoff, A. Geospatial web mining for emergency management. In *Proceedings of Sixth International Conference on Geographic Information Science, GIScience'10*, pages 1–5. Springer-Verlag, 2010.
- [78] Gabrilovich, E., Dumais, S., and Horvitz, E. Newsjunkie: Providing personalized newsfeeds via analysis of information novelty. In *Proceedings of the 13th International Conference on World Wide Web, WWW'04*, pages 482–490. ACM, 2004.
- [79] Garcia-Silva, A., Corcho, O., and Gracia, J. Associating semantics to multilingual tags in folksonomies. In *Proceedings of the 17th International Conference on Knowledge Engineering and Knowledge Management, EKAW'10*. CEUR-WS.org, 2010. URL <http://oa.upm.es/5646/>.
- [80] Gelernter, J. and Mushegian, N. Geo-parsing messages from microtext. *Transactions in GIS*, 15(6):753–773, 2011.
- [81] Go, A., Bhayani, R., and Huang, L. Twitter sentiment classification using distant supervision. In *Proceedings of the Workshop on Languages in Social Media, LSM '11*, pages 30–38. Association for Computational Linguistics, 2009.
- [82] Gockel, B., Graf, H., Pagano, A., Pescarin, S., and Eriksson, J. Vmux: An approach to user experience evaluation for virtual museums. In *Proceedings of the Second International Conference on Design, User Experience, and Usability: Design Philosophy, Methods, and Tools, DUXU'13*, pages 262–272. Springer-Verlag, 2013.
- [83] Godin, F., Debevere, P., Mannens, E., De Neve, W., and Van de Walle, R. Leveraging existing tools for named entity recognition in microposts. In *Proceedings of the Third Workshop on Making Sense of Microposts, MSM2013*, pages 36–39. CEUR-WS.org, 2013.

-
-
- [84] Gonzalez, R., Cuevas, R., Cuevas, A., and Guerrero, C. Where are my followers? understanding the locality effect in twitter. *CoRR*, abs/1105.3682, 2011.
- [85] Goolsby, R. Lifting elephants: Twitter and blogging in global perspective. In *Social Computing and Behavioral Modeling*, pages 1–6. Springer-Verlag, 2009.
- [86] Gosier, J. Swift River 2011. Online, 2011. URL <http://www.slideshare.net/Ushahidi/swiftriver-2011-overview>. Accessed: 01.04.2013.
- [87] Greater New Orleans Community Data Center (GNOCDC). Population and Loss of Children across the New Orleans Metro Area. Online, 2009. URL <http://www.gnocdc.org/repopulation>. Accessed: 01.04.2013.
- [88] Grebner, O., Bruchmann, M., Guckelsberger, C., Probst, F., and Schulz, A. Reporting and managing incidents, Patent, US 13/535,384, 2014.
- [89] Griffiths, T. L. and Steyvers, M. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(1):5228–5235, 2004.
- [90] Gu, H., Xie, X., Lv, Q., Ruan, Y., and Shang, L. Etree: Effective and efficient event modeling for real-time online social media networks. In *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT’11*, pages 300–307. IEEE Computer Society, 2011. doi: 10.1109/WI-IAT.2011.126.
- [91] Gundecha, P. and Liu, H. Mining social media: A brief introduction. *INFORMS*, 9:1–17, 2012.
- [92] Hale, S. A., Gaffney, D., and Graham, M. Where in the world are you? geolocation and language identification in twitter. In *Proceedings of the sixth International Conference on Weblogs and Social Media, ICWSM’12*, pages 518–521. AAAI Press, 2012.
- [93] Han, B., Cook, P., and Baldwin, T. A stacking-based approach to twitter user geolocation prediction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL13*, pages 7–12. Association for Computer Linguistics, 2013.
- [94] Han, B., Cook, P., and Baldwin, T. A stacking-based approach to twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, pages 451–500, 2014. to appear.
- [95] Hecht, B., Hong, L., Suh, B., and Chi, E. H. Tweets from justin beiber’s heart : The dynamics of the ”location” field in user profiles. In *Proceedings of the*



- 2011 annual conference on Human factors in computing systems, CHI '11, pages 237–246. ACM, 2011.
- [96] Heim, P. and Thom, D. Semsor: Combining social and semantic web to support the analysis of emergency situations. In *Proceedings of the 2nd Workshop on Semantic Models for Adaptive Interactive Systems, SEMAIS'11*. CEUR-WS.org, 2011.
- [97] Herald-Tribune. IbisEye: Your 2012 Hurricane source. Online, 2012. URL <http://ibiseye.com/>. Accessed: 01.04.2013.
- [98] Hienert, D., Wegener, D., and Paulheim, H. Automatic classification and relationship extraction for multi-lingual and multi-granular events from wikipedia. In *Detection, Representation, and Exploitation of Events in the Semantic Web, DeRiVE'12*, pages 1–10. CEUR-WS.org, 2012.
- [99] Hoi, S. C. H., Jin, R., and Lyu, M. R. Large-scale text categorization by batch mode active learning. In *Proceedings of the 15th international conference on World Wide Web, WWW '06*, pages 633–642. ACM, 2006.
- [100] Hong, L., Ahmed, A., Gurumurthy, S., Smola, A. J., and Tsioutsoulis, K. Discovering geographical topics in the twitter stream. In *Proceedings of the 21st international conference on World Wide Web, WWW '12*, pages 769–778. ACM, 2012.
- [101] Hsu, C.-W., Chang, C.-C., and Lin, C.-J. A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University, 2003. URL <http://www.csie.ntu.edu.tw/~cjlin/papers.html>.
- [102] Hu, X., Tang, J., Gao, H., and Liu, H. Actnet: Active learning for networked texts in microblogging. In *Proceedings of the 13th SIAM International Conference on Data Mining, SDM'13*, pages 306–314. SIAM, 2013.
- [103] Hua, T., Chen, F., Zhao, L., Lu, C.-T., and Ramakrishnan, N. Sted: Semi-supervised targeted-interest event detection in twitter. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '13*, pages 1466–1469. ACM, 2013.
- [104] Huang, S.-J., Jin, R., and Zhou, Z.-H. Active learning by querying informative and representative examples. In *Twenty-Fourth Annual Conference on Neural Information Processing Systems, NIPS'10*, pages 892–900. Springer-Verlag, 2010.
- [105] Hurlock, J. and Wilson, M. L. Searching twitter: Separating the tweet from the chaff. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, ICWSM'11*. AAAI Press, 2011.

-
-
- [106] Ikawa, Y., Enoki, M., and Tatsubori, M. User location inference using microblog messages. In *Proceedings of the 21st international conference companion on World Wide Web, WWW '12*, pages 687–690. ACM, 2012.
- [107] Ipeirotis, P. G., Provost, F. J., Sheng, V. S., and Wang, J. Repeated labeling using multiple noisy labelers. *Journal of Data Mining and Knowledge Discovery*, 28(2):402–441, 2014. URL <http://dblp.uni-trier.de/db/journals/datamine/datamine28.html#IpeirotisPSW14>.
- [108] Ishikawa, S., Arakawa, Y., Tagashira, S., and Fukuda, A. Hot topic detection in local areas using twitter and wikipedia. In *Proceedings of ARCS Workshops, ARCS'12*, pages 1–5. IEEE Computer Society, 2012.
- [109] Jadhav, A., Wang, W., Mutharaju, R., and Anantharam, P. Twitris: Socially influenced browsing. In *Semantic Web Challenge 2009, ISWC'09*. ACM, 2009.
- [110] Jain, A. K., Murty, M. N., and Flynn, P. J. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, 1999.
- [111] Japkowicz, N. and Shah, M. *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press, 2011.
- [112] Jiang, M. and McGill, W. L. Human-centered sensing for crisis response and management analysis campaigns. In *Proceedings of the 7th International ISCRAM Conference, ISCRAM'10*. ISCRAM, 2010.
- [113] Joachims, T. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning, ECML '98*, pages 137–142. Springer-Verlag, 1998.
- [114] JRC European Commission. MediSys. Online, 2012. URL <http://medusa.jrc.it/medisys>. Accessed: 01.04.2013.
- [115] Jurgens, D. That's what friends are for: Inferring location in online social media platforms based on social relationships. In *Proceedings of Seventh International AAAI Conference on Weblogs and Social Media, ICWSM'13*. AAAI Press, 2013.
- [116] Kang, J., Ryu, K. R., and Kwon, H. C. Using cluster-based sampling to select initial training set for active learning in text classification. In *Proceedings of 8th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, PAKDD'04*, pages 384–388. Springer-Verlag, 2004.
- [117] Kantardzic, M. *Data Mining: Concepts, Models, Methods and Algorithms*. John Wiley & Sons, Inc., 2011.
- [118] Kaplan, A. M. and Haenlein, M. Users of the world, unite! the challenges and opportunities of social media. *Business Horizons*, 53(1):59–68, 2010.

-
-
- [119] Karimi, S., Yin, J., and Paris, C. Classifying microblogs for disasters. In *Proceedings of the 18th Australasian Document Computing Symposium, ADCS '13*, pages 26–33. ACM, 2013.
- [120] Karypis, G., Aggarwal, R., Kumar, V., and Shekhar, S. Multilevel hypergraph partitioning: Application in vlsi domain. In *Proceedings of the 34th Annual Design Automation Conference, DAC '97*, pages 526–529. ACM, 1997.
- [121] Kinsella, S. and Murdock, V. I'm eating a sandwich in glasgow: modeling locations with tweets. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents, SMUC '11*, pages 61–68. ACM, 2011.
- [122] Kleinberg, J. Bursty and hierarchical structure in streams. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '02*, pages 91–101. ACM, 2002.
- [123] Krishnamurthy, B. and Arlitt, M. A few chirps about twitter. In *Proceedings of the first workshop on Online social networks, WOSN '08*, pages 19–24. ACM, 2006.
- [124] Kulshrestha, J., Kooti, F., Nikraves, A., and Gummadi, K. Geographic dissection of the twitter network. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media, ICWSM'12*. AAAI Press, 2012.
- [125] Kwak, H., Lee, C., Park, H., and Moon, S. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web, WWW '10*, pages 591–600. ACM, 2010.
- [126] Lafferty, J. D., McCallum, A., and Pereira, F. C. N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01*, pages 282–289. Morgan Kaufmann Publishers Inc., 2001.
- [127] Lane, N. D., Eisenman, S. B., Musolesi, M., Miluzzo, E., and Campbell, A. T. Urban sensing systems: Opportunistic or participatory? In *Proceedings of the 9th Workshop on Mobile Computing Systems and Applications, HotMobile '08*, pages 11–16. ACM, 2008.
- [128] Laugwitz, B., Held, T., and Schrepp, M. Construction and evaluation of a user experience questionnaire. In *Proceedings of the 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society on HCI and Usability for Education and Work, USAB '08*, pages 63–76. Springer-Verlag, 2008.
- [129] Le-Phuoc, D., Quoc, H., and Parreira, J. The linked sensor middleware – connecting the real world and the semantic web. In *Semantic Web Challenge 2011, ISWC 2011*. Springer-Verlag, 2011.

-
-
- [130] Le-Phuoc, D., Parreira, J. X., Hausenblas, M., Han, Y., and Hauswirth, M. Live linked open sensor database. In *Proceedings of the 6th International Conference on Semantic Systems, I-SEMANTICS '10*, pages 46:1–46:4. ACM, 2010.
- [131] Lee, R. and Sumiya, K. Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection. In *Proceedings of the 2Nd ACM SIGSPATIAL International Workshop on Location Based Social Networks, LBSN '10*, pages 1–10. ACM, 2010.
- [132] Leidner, J. L. Toponym resolution in text : "which sheffield is it?". In *Proceedings of the the 27th Annual International ACM SIGIR Conference, SIGIR '04*, pages 602–602. ACM, 2004.
- [133] Lewis, D. D. and Catlett, J. Heterogeneous uncertainty sampling for supervised learning. In *Proceedings of the 11th International Conference on Machine Learning, ICML'94*, pages 148–156. Morgan Kaufmann Publishers Inc., 1994. URL citeseer.nj.nec.com/135290.html.
- [134] Lewis, D. D. and Gale, W. A. A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference, SIGIR '94*, pages 3–12. ACM, 1994.
- [135] Ley, B., Pipek, V., Reuter, C., and Wiedenhoefer, T. Supporting inter-organizational situation assessment in crisis management. In *Proceedings of the 9th International ISCRAM Conference, ISCRAM'12*, pages 1–10. ISCRAM, 2012.
- [136] Li, C., Sun, A., and Datta, A. Twevent: Segment-based event detection from tweets. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM '12*, pages 155–164. ACM, 2012.
- [137] Li, R., Lei, K. H., Khadiwala, R., and Chang, K. C.-C. Tedas: A twitter-based event detection and analysis system. In *Proceedings of the 28th International Conference on Data Engineering, ICDE'12*, pages 1273–1276. IEEE Computer Society, 2012.
- [138] Li, X. and Croft, W. B. Time-based language models. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management, CIKM '03*, pages 469–475. ACM, 2003.
- [139] Lieberman, M. D., Samet, H., and Sankaranarayanan, J. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *Proceedings of the 26th International Conference on Data Engineering, ICDE 2010*, pages 201–212. IEEE Computer Society, 2010.

-
-
- [140] Lin, J., Snow, R., and Morgan, W. Smoothing techniques for adaptive on-line language models: Topic tracking in tweet streams. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '11*, pages 422–429. ACM, 2011.
- [141] Long, R., Wang, H., Chen, Y., Jin, O., and Yu, Y. Towards effective event detection, tracking and summarization on microblog data. In *Proceedings of the 12th International Conference on Web-age Information Management, WAIM'11*, pages 652–663. Springer-Verlag, 2011.
- [142] Lui, S. and Palen, L. The new cartographers: Crisis map mashups and the emergence of neogeographic practice. *Cartography and Geographic Information Science*, 37(1):69–90, 2010.
- [143] MacEachren, A. M., Jaiswal, A., Robinson, A. C., Pezanowski, S., Savelyev, A., Mitra, P., Zhang, X., and Blanford, J. Senseplace2: Geotwitter analytics support for situational awareness. In *Proceedings of 2011 IEEE Conference on Visual Analytics Science and Technology, VAST '11*, pages 181–190. IEEE Computer Society, 2011.
- [144] Mahmud, J., Nichols, J., and Drews, C. Where is this tweet from? inferring home locations of twitter users. In *Proceedings of Sixth International AAAI Conference on Weblogs and Social Media, ICWSM'12*. AAAI Press, 2012.
- [145] Manning, C. D., Raghavan, P., and Schütze, H. *An Introduction to Information Retrieval*. Cambridge University Press, 2009.
- [146] Marcus, A., Bernstein, M. S., Badar, O., Karger, D. R., Madden, S., and Miller, R. C. Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI'11*, pages 227–236. ACM, 2011.
- [147] Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., and Gómez-Berbás, J. M. Named entity recognition: Fallacies, challenges and opportunities. *Journal of Computer Standards and Interfaces*, 32(5):482–489, 2012.
- [148] Massoudi, K., Tsagkias, M., de Rijke, M., and Weerkamp, W. Incorporating query expansion and quality indicators in searching microblog posts. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval, ECIR'11*, pages 362–367. Springer-Verlag, 2011.
- [149] Mathioudakis, M. and Koudas, N. Twittermonitor: Trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data, SIGMOD '10*, pages 1155–1158. ACM, 2010.

-
-
- [150] Mazumdar, D., Lanfranchi, V., Cano, A., and F., C. Visualising topical sentiment and influence in social media. In *Proceedings of Social Media and Linked Data for Emergency Response (SMILE) Workshop, ESWC'13*. Springer-Verlag, 2013.
- [151] Mccallum, A. and Nigam, K. A comparison of event models for naive bayes text classification. Technical report ws-98-05, AAAI Press, 1998.
- [152] Mcgee, J., Caverlee, J., and Cheng, Z. A geographic study of tie strength in social media. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011*, pages 2333–2336. ACM, 2011.
- [153] McMinn, A. J., Moshfeghi, Y., and Jose, J. M. Building a large-scale corpus for evaluating event detection on twitter. In *Proceedings of the 22Nd ACM International Conference on Conference on Information and Knowledge Management, CIKM '13*, pages 409–418. ACM, 2013.
- [154] Mencía, E. L., Holthausen, S., Schulz, A., and Janssen, F. Using data mining on linked open data for analyzing e-procurement information. In *Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases ECML/PKDD 2013, Data Mining on Linked Data Workshop*. CEUR-WS.org, 2013.
- [155] Mendes, P. N., Jakob, M., García-Silva, A., and Bizer, C. Dbpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems, I-Semantics'11*. ACM, 2011.
- [156] Metzler, D., Cai, C., and Hovy, E. Structured event retrieval over microblog archives. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT '12*, pages 646–655. Association for Computational Linguistics, 2012.
- [157] Muñoz García, O., García-Silva, A., Corcho, O., de la Higuera Hernández, M., and Navarro, C. Identifying topics in social media posts using dbpedia. In *Proceedings of the NEM Summit*, pages 81–86. NEM, 2011.
- [158] Nadeau, D. and Sekine, S. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26, 2007. John Benjamins Publishing Company.
- [159] National Association of Radio Distress-Signalling and Infocommunications. RSOE EDIS. Online, 2012. URL <http://hisz.rsoe.hu/>. Accessed: 01.04.2013.
- [160] Nelder, A. and Mead, R. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313, 1965.

-
-
- [161] Newman, M. E. J. Detecting community structure in networks. *The European Physical Journal B*, 38(2):321–330, 2004.
- [162] Nguyen, H. T. and Smeulders, A. Active learning using pre-clustering. In *Proceedings of the twenty-first international conference on Machine learning, ICML '04*, pages 79–86. ACM, 2004.
- [163] Norman, D. A. *The Design of Everyday Things*. Basic Books, 2002.
- [164] Oh, O., Agrawal, M., and Rao, H. R. Information control and terrorism: Tracking the mumbai terrorist attack through twitter. *Information Systems Frontiers*, 13(1):33–43, 2011.
- [165] Okolloh, O. Ushahidi, or 'testimony': Web 2.0 tools for crowdsourcing crisis information. *Participatory Learning and Action*, 59(1):65–70, 2009.
- [166] Olsson, F. A literature survey of active machine learning in the context of natural language processing. Technical report, Swedish Institute of Computer Science, 2009. URL <http://soda.swedish-ict.se/3600/1/SICS-T--2009-06--SE.pdf>.
- [167] Packer, H. S., Samangooui, S., Hare, J. S., Gibbins, N., and Lewis, P. H. Event detection using twitter and structured semantic query expansion. In *Proceedings of the 1st International Workshop on Multimodal Crowd Sensing, Crowd-Sens '12*, pages 7–14. ACM, 2012.
- [168] Page, L., Brin, S., Motwani, R., and Winograd, T. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford University, 1998. URL [/brokenurl#http://publication.wilsonwong.me/load.php?id=233281827](http://publication.wilsonwong.me/load.php?id=233281827).
- [169] Paradesi, S. Geotagging tweets using their content. In *Proceedings of the Twenty-fourth International Florida Artificial Intelligence Research Society, FLAIRS '11*, pages 355–356. AAAI Press, 2011.
- [170] Parikh, R. and Karlapalem, K. Et: Events from tweets. In *Proceedings of the 22nd International Conference on World Wide Web Companion, WWW'13*, pages 613–620. International World Wide Web Conferences Steering Committee, 2013.
- [171] Paulheim, H. Exploiting linked open data as background knowledge in data mining. In *Proceedings of Data Mining on Linked Data Workshop, DMO LD'13*. CEUR-WS.org, 2013.
- [172] Petasis, G., Cucchiarelli, A., Velardi, P., Paliouras, G., Karkaletsis, V., and Spyropoulos, C. D. Automatic adaptation of proper noun dictionaries through cooperation of machine learning and probabilistic methods. In *Proceedings of*

-
- the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '00*, pages 128–135. ACM, 2000.
- [173] Phuvipadawat, S. and Murata, T. Breaking news detection and tracking in twitter. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT '10*, pages 120–123. IEEE Computer Society, 2010.
- [174] Platt, J. C. *Fast Training of Support Vector Machines using Sequential Minimal Optimization*, pages 185–208. MIT Press, 1999.
- [175] Popescu, A. and Grefenstette, G. Mining user home location and gender from flickr tags. In *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM'10*, pages 307–310. AAAI Press, 2010.
- [176] Popescu, A.-M., Pennacchiotti, M., and Paranjpe, D. Extracting events and event descriptions from twitter. In *Proceedings of the 20th International Conference Companion on World Wide Web, WWW '11*, pages 105–106. ACM, 2011.
- [177] Pozdnoukhov, A. and Kaiser, C. Space-time dynamics of topics in streaming text. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks, LBSN '11*, pages 1–8. ACM, 2011.
- [178] Puras, J. C. and Iglesias, C. A. Disasters 2.0: Application of web 2.0 technologies in emergency situations. In *Proceedings of the 6th International ISCRAM Conference, ISCRAM'09*. ISCRAM, 2009.
- [179] Raghavan, H., Madani, O., Jones, R., and Kaelbling, P. Active learning with feedback on both features and instances. *Journal of Machine Learning Research*, 7:1655–1686, 2006.
- [180] Raina, R., Ng, A. Y., and Koller, D. Constructing informative priors using transfer learning. In *Proc of the 23rd International Conference on Machine Learning, ICML '06*, pages 713–720. ACM, 2006.
- [181] Rajaraman, A. and Ullman, J. D. *Mining of Massive Datasets*. Cambridge University Press, 2011.
- [182] Report, P What you see about floods. Online, 2012. URL <http://pakreport.org/flood2010/>. Accessed: 01.12.2012.
- [183] Ritter, A., Clark, S., Mausam, and Etzioni, O. Named entity recognition in tweets: An experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11*, pages 1524–1534. Association for Computational Linguistics, 2011.

-
-
- [184] Ritter, A., Mausam, Etzioni, O., and Clark, S. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12*, pages 1104–1112. ACM, 2012.
- [185] Rizzo, G. and Troncy, R. Nerd: a framework for evaluating named entity recognition tools in the web of data. In *Proceedings at the 13th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2012*, pages 73–76. Association for Computational Linguistics, 2012.
- [186] Robert Power, D. R., Bella Robinson. Finding fires with twitter. In *Australasian Language Technology Association Workshop*, pages 80–89. Association for Computational Linguistics, 2013.
- [187] Rodrigues, E., Assuncao, R., Pappa, G., Miranda, R., and Meira, W. Uncovering the location of twitter users. In *2013 Brazilian Conference on Intelligent Systems, BRACIS*, pages 237–241. IEEE Computer Society, 2013.
- [188] Romero, D. M., Meeder, B., and Kleinberg, J. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th International Conference on World Wide Web, WWW '11*, pages 695–704. ACM, 2011.
- [189] Ruiz, E. J., Hristidis, V., Castillo, C., Gionis, A., and Jaimes, A. Correlating financial time series with micro-blogging activity. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, WSDM'12*, pages 513–522. ACM, 2012.
- [190] Sadilek, A., Kautz, H., and Bigham, J. P. Finding your friends and following them to where you are. In *Proceedings of the fifth ACM international conference on Web search and data mining, WSDM'12*, pages 723–732. ACM, 2012.
- [191] Saif, H., He, Y., and Alani, H. Semantic sentiment analysis of twitter. In *Proceedings of the 11th International Conference on The Semantic Web, ISWC'12*, pages 508–524. Springer-Verlag, 2012.
- [192] Sakaki, T. and Okazaki, M. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World Wide Web, WWW '10*, pages 851–860. ACM, 2010.
- [193] Sankaranarayanan, J., Samet, H., Teitler, B. E., Lieberman, M. D., and Sperling, J. Twitterstand: News in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS '09*, pages 42–51. ACM, 2009.

-
-
- [194] Schulz, A., Paulheim, H., and Probst, F. Crisis information management in the web 3.0 age. In *Proceedings of the 9th International Conference on Information Systems for Crisis Response and Management, ISCRAM'12*. ISCRAM, 2012.
- [195] Schulz, A. and Paulheim, H. Combining government and linked open data in emergency management. In *AI Mashup Challenge 2012 colocated with 9th Extended Semantic Web Conference (ESWC 2012)*, 2012.
- [196] Schulz, A. and Paulheim, H. Mashups for the emergency management domain. In *Semantic Mashups*, pages 237–260. Springer-Verlag, 2013.
- [197] Schulz, A. and Ristoski, P. The car that hit the burning house: Understanding small scale incident related information in microblogs. In *Proceedings of the 2nd When the City Meets the Citizen Workshop (WCMCW) at ICWSM'13*. AAAI Press, 2013. URL <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6196>.
- [198] Schulz, A., Döweling, S., and Probst, F. Integrating process modeling and linked open data to improve decision making in disaster management. In *Proceedings of the CSCW 2012 Workshop on Collaboration and Crisis Informatics*, pages 16–23. International Institute for Socio-Informatics, 2012.
- [199] Schulz, A., Ortmann, J., and Probst, F. Getting user-generated content structured: Overcoming information overload in emergency management. In *Proceedings of 2012 IEEE Global Humanitarian Technology Conference, GHTC'12*, pages 1–10. IEEE Computer Society, 2012.
- [200] Schulz, A., Hadjakos, A., Paulheim, H., Nachtwey, J., , and Mühlhäuser, M. A multi-indicator approach for geolocalization of tweets. In *Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM'13*. AAAI Press, 2013. URL <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6063>.
- [201] Schulz, A., Ristoski, P., and Paulheim, H. I see a car crash: Real-time detection of small scale incidents in microblogs. In *The Semantic Web: ESWC 2013 Satellite Events*, volume 7955 of *Lecture Notes in Computer Science*, pages 22–33. Springer-Verlag, 2013.
- [202] Schulz, A., Thanh, T. D., Paulheim, H., and Schweizer, I. A fine-grained sentiment analysis approach for detecting crisis related microposts. In *Komplexe Notsituationen schnell meistern - Die ISCRAM Konferenz 2013 zum Krisenmanagement, ISCRAM'13*, pages 846–851. ISCRAM, 2013.
- [203] Schulz, A., Mencía, E. L., Dang, T. T., and Schmidt, B. Evaluating multi-label classification of incident-related tweets. In *Proceedings of the WWW'14 Workshop on Microposts*. CEUR-WS.org, 2014.

-
- [204] Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Lucas, R. E., Agrawal, M., Park, G. J., Lakshmikanth, S. K., Jha, S., Seligman, M. E. P., and Ungar, L. H. Characterizing geographic variation in well-being using tweets. In *Proceedings of the Seventh International AAI Conference on Weblogs and Social Media, ICWSM'13*. AAAI Press, 2013.
- [205] Securities and Commission, E. Form s-1 - securities and exchange commission - twitter, inc. Online, 2013. URL <http://www.sec.gov/Archives/edgar/data/1418091/000119312513390321/d564001ds1.htm>. Accessed: 20.03.2014.
- [206] Sequeda, J. F. and Corcho, O. Linked stream data: A position paper. In *Proceedings of the 3rd International Workshop on Semantic Sensor Networks, SSN'10*, pages 148–157. CEUR-WS.org, 2009.
- [207] Serdyukov, P., Murdock, V., and van Zwol, R. Placing flickr photos on a map. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, SIGIR '09*, pages 484–491. ACM, 2009.
- [208] Settles, B. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2012.
- [209] Seung, H. S., Oppen, M., and Sompolinsky, H. Query by committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory, COLT '92*, pages 287–294. ACM, 1992.
- [210] Shannon, C. E. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 1948.
- [211] Shen, D., Zhang, J., Su, J., Zhou, G., and Tan, C.-L. Multi-criteria-based active learning for named entity recognition. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, ACL'04*. Association for Computational Linguistics, 2004.
- [212] Sheng, V. S., Provost, F., and Ipeirotis, P. G. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '08*, pages 614–622. ACM, 2008.
- [213] Signorini, A., Segre, A. M., and Polgreen, P. M. The use of twitter to track levels of disease activity and public concern in the u.s. during the influenza a h1n1 pandemic. *PLoS ONE*, 6(5), 2011.
- [214] Sigurbjörnsson, B. and Van Zwol, R. Flickr tag recommendation based on collective knowledge. In *Proceeding of the 17th international conference on World Wide Web, WWW '08*, pages 327–336. ACM, 2008.

-
-
- [215] Silverman, B. W. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.
- [216] Smith, D. A. and Crane, G. Disambiguating geographic names in a historical digital library. In *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries, ECDL '01*, pages 127–136. Springer-Verlag, 2001.
- [217] Song, Y., Wang, H., Wang, Z., Li, H., and Chen, W. Short text conceptualization using a probabilistic knowledgebase. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, IJCAI'11*, pages 2330–2336. AAAI Press, 2011.
- [218] Srivastava, M., Abdelzaher, T., and Szymanski, B. Human-centric sensing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 370(1958):176–197, 2012.
- [219] Starbird, K., Palen, L., Hughes, A. L., and Vieweg, S. Chatter on the red: what hazards threat reveals about the social life of microblogged information. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work, CSCW '10*, pages 241–250. ACM, 2010.
- [220] Steiner, T., Verborgh, R., Gabarró Vallés, J., and Van de Walle, R. Adding meaning to social network microposts via multiple named entity disambiguation APIs and tracking their data provenance. *International Journal of Computer Information Systems and Industrial Management*, 5:69–78, 2013.
- [221] Strötgen, J. and Gertz, M. Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, 47(2):269–298, 2012.
- [222] Sultanik, E. A. and Fink, C. Rapid geotagging and disambiguation of social media text via an indexed gazetteer. In *Proceedings of the Conference on Information Systems for Crisis Response and Management, ISCRAM '12*. ISCRAM, 2012.
- [223] Takahashi, T., Abe, S., and Igata, N. *Human-Computer Interaction. Towards Mobile and Intelligent Interaction Environments.*, volume 6763 of *Lecture Notes in Computer Science*, chapter Can Twitter Be an Alternative of Real-World Sensors?, pages 240–249. Springer-Verlag, 2011.
- [224] Takhteyev, Y., Gruz, A., and Wellman, B. Geography of twitter networks. *Social Networks*, 34(1):73–81, 2012.
- [225] Tang, M., Luo, X., and Roukos, S. Active learning for statistical natural language parsing. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02*, pages 120–127. Association for Computational Linguistics, 2002.

-
- [226] Tapia, A., Bajpai, K., Jansen, B., and Yen, J. Seeking the trustworthy tweet: Can microblogged data fit the information needs of disaster response and humanitarian relief organizations? In *Proceedings of the 8th International Conference on Information Systems for Crisis Response and Management, ISCRAM'11*. ISCRAM, 2011.
- [227] Terpstra, T., de Vries, A., Stronkman, R., and Paradies, G. Towards a realtime twitter analysis during crises for operational crisis management. In *Proceedings of 9th International Conference on Information Systems for Crisis Response and Management, ISCRAM'12*. ISCRAM, 2012.
- [228] The World Factbook 2013-14, Central Intelligence Agency, Washington, DC. Regional and world maps. Online, 2014. URL <http://upload.wikimedia.org/wikipedia/commons/e/e8/Timezoneswest.PNG>. Accessed: 01.02.2014.
- [229] Thongsuk, C., Haruechaiyasak, C., and Meesad, P. Classifying business types from twitter posts using active learning. In *Proceedings of 10th International Conference on Innovative Internet Community Services, I2CS*, pages 180–189. Springer-Verlag, 2010.
- [230] Times, L. A. Los angeles fire map. Online, 2012. URL <http://www.latimes.com/news/local/la-me-la-fire-map-html,0,7464337.Htmlstory/>. Accessed: 01.13.2013.
- [231] Tong, S. and Koller, D. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2:45–66, 2002.
- [232] Twitter, I. Twitter reports fourth quarter and fiscal year 2013 results. Online, 2014. URL http://files.shareholder.com/downloads/AMDA-2F526X/3037384043x0x723604/deb9d853-90d7-4ff4-b1b7-849bb2d42890/TWTR_News_2014_2_5_Financial_releases.pdf. Accessed: 04.03.2014.
- [233] University of Fortaleza. Wikicrimes - mapping crimes collaboratively. Online, 2012. URL <http://www.wikicrimes.org>. Accessed: 01.04.2013.
- [234] Verhagen, M. and Pustejovsky, J. Temporal processing with the tarsqi toolkit. In *Proceedings of 22nd International Conference on Computational Linguistics: Demonstration Papers*, pages 189–192. Association for Computational Linguistics, 2008.
- [235] Verhagen, M., Saurí, R., Caselli, T., and Pustejovsky, J. Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval '10*, pages 57–62. Association for Computational Linguistics, 2010.

-
-
- [236] Vieweg, S., Hughes, A., Starbird, K., and Palen, L. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the 28th international conference on Human factors in computing systems, CHI'10*, pages 1079–1088. ACM, 2010.
- [237] Wallace, B. C., Small, K., Brodley, C. E., and Trikalinos, T. A. Who should label what? instance allocation in multiple expert active learning. In *Proceedings of the SIAM Conference on Data Mining, SDM'11*, pages 176–187. Omnipress, 2011.
- [238] Walther, M. and Kaisser, M. Geo-spatial event detection in the twitter stream. In *Proceedings of the 35th European Conference on Advances in Information Retrieval, ECIR'13*, pages 356–367. Springer-Verlag, 2013.
- [239] Wang, X., Xu, M., Ren, Y., Xu, J., Zhang, H., and Zheng, N. A location inferring model based on tweets and bilateral follow friends. *Journal of Computers*, 9 (2), 2014.
- [240] Wanichayapong, N., Pruthipunyaskul, W., Pattara-Atikom, W., and Chaovalit, P. Social-based traffic information extraction and classification. In *Proceedings of the 11th International Conference on ITS Telecommunications, ITST'11*, pages 107–112. IEEE Computer Society, 2011.
- [241] Watanabe, K., Ochi, M., Okabe, M., and Onai, R. Jasmine: A real-time local-event detection system based on geolocation information propagated to microblogs. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management, CIKM '11*, pages 2541–2544. ACM, 2011.
- [242] Weber, I., Garimella, V. R. K., and Batayneh, A. Secular vs. islamist polarization in egypt on twitter. In *roceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM'13*, pages 290–297. ACM, 2013.
- [243] Weng, J. and Lee, B.-S. Event detection in twitter. In *Proceedings of the Fifth International Conference on Weblogs and Social Media, ICWSM'11*. AAAI Press, 2011. URL <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2767>.
- [244] Williams, C. K. I. *Learning in Graphical Models*, volume 89 of *NATO ASI Series*, chapter Prediction With Gaussian Processes: From Linear Regression To Linear Prediction And Beyond, pages 599–621. Springer-Verlag, 1997.
- [245] Witten, I. H. and Frank, E. *Data mining: practical machine learning tools and techniques*. Elsevier, 2005.

-
-
- [246] Woodruff, A. G. and Plaunt, C. Gipsy : Automated geographic indexing of text documents previous work in georeferencing of text documents. *Journal of the American Society for Information Science*, 45(9):645–655, 1994.
- [247] Xie, K., Xia, C., Grinberg, N., Schwartz, R., and Naaman, M. Robust detection of hyper-local events from geotagged social media data. In *Proceedings of the Thirteenth International Workshop on Multimedia Data Mining, MDMKDD '13*, pages 2:1–2:9. ACM, 2013.
- [248] Xu, T. and Oard, D. W. Wikipedia-based topic clustering for microblogs. In *Proceedings of the American Society for Information Science and Technology, ASIST'11*, pages 1–10. Wiley, 2011.
- [249] Xu, Z., Yu, K., Tresp, V., Xu, X., and Wang, J. Representative sampling for text classification using support vector machines. In *Proceedings of the 25th European conference on IR research, ECIR '03*, pages 393–407. ACM, 2003.
- [250] Yang, Y., Carbonell, J., Brown, R., Pierce, T., Archibald, B., and Liu, X. Learning approaches for detecting and tracking news events. *IEEE Intelligent Systems*, 14(4):32–43, 1999.
- [251] Yang, Y. and Pedersen, J. O. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning, ICML '97*, pages 412–420. Morgan Kaufmann Publishers Inc., 1997.
- [252] Yang, Y., Pierce, T., and Carbonell, J. A study of retrospective and on-line event detection. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '98*, pages 28–36. ACM, 1998.
- [253] Yuen, M.-C., King, I., and Leung, K.-S. A survey of crowdsourcing systems. In *Proceedings of the third International Conference on Social Computing, Social-Com*, pages 766–773. IEEE Computer Society, 2011.
- [254] Zang, N., Rosson, M., and Nasser, V. Mashups: who? what? why? In *Extended Abstracts on Human Factors in Computing Systems, CHI'08*, pages 3171–3176. ACM, 2008.
- [255] Zhao, L., Sukthankar, G., and Sukthankar, R. Incremental relabeling for active learning with noisy crowdsourced annotations. In *Proceedings of the third International Conference on Social Computing, SocialCom*, pages 728–733. IEEE Computer Society, 2011.
- [256] Zhu, J., Wang, H., Yao, T., and Tsou, B. K. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In

Proc. of the 22nd International Conference on Computational Linguistics, COLING '08, pages 1137–1144. Association for Computational Linguistics, 2008.

- [257] Zong, W., Wu, D., Sun, A., Lim, E.-P., and Goh, D. H.-L. On assigning place names to geography related web pages. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries, JCDL '05*, pages 354–362. ACM, 2005.



List of Figures

1.	Overview of the connections of the chapters in this dissertation.	8
2.	Overview of the framework for small-scale incident detection	15
3.	Collection and filtering as the first step in the framework.	17
4.	Overview of the connections of Part II to the chapters in this dissertation.	26
5.	Automatic preprocessing as a second step of the overall framework. . .	27
6.	Extended preprocessing pipeline for social media texts based on the steps proposed by Manning et al. [145].	29
7.	Automatic preprocessing as a second step of the overall pipeline.	41
8.	Example for spatial indicators in tweets and user profiles.	44
9.	Countries covered by time zone entry "Eastern Time (USA&Canada)". (Picture adapted based on [228])	45
10.	Countries covered by UTC offset of "UTC-05:00". (Picture adapted based on [228])	46
11.	The height profile is determined by stacking the three-dimensional polygon shapes over each other.	52
12.	Example pipeline for our approach: Spatial indicators are identified based on the methods described. The results are either a pair of coordinates (<i>lat,lon</i>) or a set of coordinates and quality measures. The coordinates are mapped to the corresponding polygons. Then the external quality measures are applied before conducting the stacking. As a result we estimate the location of the tweet with a confidence value.	53
13.	Example pipeline for our approach for street-level geolocalization of incident-related tweets.	58
14.	Overview of evaluation results for LBS and SP. The y-axis of each chart provides the percentages of the tweets within spatial distance. The x-axis shows the spatial distance in km.	62
15.	Overview of evaluation results for applying the approaches on the location field. The y-axis of each chart provides the percentages of the tweets within spatial distance. The x-axis shows the spatial distance in km.	64
16.	Overview of evaluation results for TZ, WS-1, and WS-2. The y-axis of each chart provides the percentages of the tweets within spatial distance. The x-axis shows the spatial distance in km.	65

17.	Overview of the overall results with external quality measures (bottom right). The y-axis of the chart provides the percentages of the tweets within spatial distance. The x-axis shows the spatial distance in km.	66
18.	Overview of the overall results for estimating the user's residence. The y-axis of the chart provides the percentages of the tweets within spatial distance. The x-axis shows the spatial distance in km.	68
19.	Overview of the overall results for estimating the focus of incident-related tweets. The y-axis of the chart provides the percentages of the tweets within spatial distance. The x-axis shows the spatial distance in km.	69
20.	Overview of the connections of Part III to the chapters in this dissertation.	75
21.	The human-based classification and aggregation step in the framework.	79
22.	Screenshot of the Ushahidi platform for tracking events in the Gaza Strip.	88
23.	Example process for human-based classification.	90
24.	Participatory human-centered sensing and human-based classification integrated into the overall framework.	92
25.	Screenshot of a prototype leveraging participatory human-centered sensing and human-based classification for emergency management. .	95
26.	The results of the user experience questionnaire for Task 1.	98
27.	The results of the user experience questionnaire for Task 2.	99
28.	Example of automatically mapping incident reports to questions based on named entities.	106
29.	Machine-based classification as a step in the framework.	107
30.	Example of decision hyperplane, margin, and support vectors for a binary linear classifier.	111
31.	Pipeline for automatic incident type classification showing semantic abstraction, feature generation, and training steps.	120
32.	RapidMiner process for extracting direct types of entities.	122
33.	Integration of classified tweets into the Incident Classifier application.	128
34.	Class distribution for 4-CLASSES and 2-CLASSES data sets.	130
35.	Learning curves for different semantic abstraction approaches compared with the baseline on 4-CLASSES data set.	149
36.	Learning curves for different semantic abstraction approaches compared with the baseline on 2-CLASSES data set.	150
37.	Machine-based aggregation as a step in the framework.	169

38.	Two examples showing the clustering of two incident reports. In the first example, the spatial, temporal, and thematic dimensions match; thus, we assume that both incident reports are about the same incident. In the second example, a difference in the temporal dimension is present; thus, both incident reports are most likely about different incidents.	170
39.	Approach for detecting incidents and clustering incident reports. . . .	184
40.	Assignment of incident types within and across incident type vocabularies.	185
41.	The four cases when the thematic extent of an existing incident matches a newly reported incident. The numbers correspond to the enumeration in the text above.	188
42.	The user interface to create rules. The vocabulary contains the German terms for the fire brigade involved in the InfoStrom project. . . .	190
43.	SID as an application for incident detection based on user-generated content.	192
44.	SID as an application for incident detection based on user-generated content.	193
45.	Distribution of the number of users and the number of tweets from different user categories by incident type in SET_1_L.	196
46.	Refinement as a step in the framework.	209
47.	Initial selection and query selection as parts of the refinement step. . .	212
48.	Example of a binary classification problem with respective decision boundaries of an SVM. Furthermore, the instances selected by uncertainty sampling and representative sampling are shown.	213
49.	Instances selected by different selection strategies for a binary classification problem.	219
50.	Evaluation results of state-of-the-art selection strategies and our approach. The graphs for every combination of annotators with noise and without noise are shown.	230
51.	Evaluation results of state-of-the-art selection strategies and our approach. The graphs for every combination of annotators with noise and without noise are shown.	231
52.	Influence of noise on our approach. The graphs for every combination of annotators with and without noise are shown.	233
53.	The graphs for one, five, and 200 annotators are shown with different combinations of batch sizes.	234



List of Tables

1.	Real-world incident types and number of extracted keywords.	22
2.	Named entity recognition rate for tweets using Spotlight.	38
3.	Evaluation results of location mention extraction approach.	38
4.	Evaluation results for temporal expression extraction.	38
5.	Overview of related approaches. Spatial indicators and techniques marked with (X) were used by the respective approaches for creating baselines or were part of the background analysis.	47
6.	Optimal external quality measures for each spatial indicator.	61
7.	Results of the individual indicator approaches (in km) and external quality measures of the indicators.	62
8.	Results of the individual indicator approaches (in km) and external quality measures of the indicators.	63
9.	Results of the overall geolocalization approach for tweets with and without external quality measures.	66
10.	Recall of individual indicator approaches on a random and unfiltered sample of the Spritzer stream.	67
11.	Results of the overall approach for estimating the user's residence. . . .	68
12.	Comparison of our approach for estimating the user's residence with related approaches.	69
13.	Results of the overall approach for estimating the focus of incident-related tweets.	69
14.	Overview of approaches that apply crowdsourcing for processing incident-related information.	85
15.	Years of work experience of user study participants.	97
16.	Results for the overall error, random error, and systematic error (μ =mean, σ =standard deviation).	104
17.	Overview of related approaches for incident type classification.	113
18.	Overview of related approaches for incident type classification with respect to the used feature groups as well as according to the use of semantic abstraction. (Named Entities = NEs)	114
19.	Overview of approaches that are related to our semantic abstraction. .	116
20.	Extracted Types and Categories for Tweet "Car crash on Interstate 90, everything is on fire".	122
21.	Features and emoticons used for EMO feature group.	126

22.	The two-class data sets for evaluating the semantic abstraction approach.	131
23.	The four-class data sets used for evaluating the semantic abstraction approach.	131
24.	Classification results when keyword-based classification is applied on 4-CLASSES.	133
25.	Classification results when keyword-based classification is applied on 2-CLASSES.	134
26.	Evaluation results for different n-gram combinations and weighting strategies on 4-CLASSES data set.	136
27.	Evaluation results for different n-gram combinations and weighting strategies on 2-CLASSES data set.	137
28.	Evaluation results for replacement strategies before n-gram generation on 4-CLASSES and 2-CLASSES data sets.	139
29.	Evaluation results for syntactic features on 4-CLASSES and 2-CLASSES data sets.	140
30.	Evaluation results for the emoticon features before n-gram generation on 4-CLASSES and 2-CLASSES data sets.	140
31.	Evaluation results for TF-IDF scores on 4-CLASSES and 2-CLASSES data sets.	142
32.	Evaluation results for abstracting temporal expressions before n-gram generation on 4-CLASSES and 2-CLASSES data sets.	143
33.	Evaluation results for location mention abstraction before n-gram generation on 4-CLASSES and 2-CLASSES data sets.	144
34.	Evaluation results for the LOD feature groups before n-gram generation on 4-CLASSES and 2-CLASSES data sets.	145
35.	Increase in F-measure for using semantic abstraction compared with a baseline comprising n-gram features after Slang and URL replacement, TF-IDF scores, and syntactic features.	147
36.	Contingency table for conducting McNemar's test.	152
37.	Evaluation of semantic abstraction when trained on Chicago and Memphis data sets. (* significant difference in error rates with $p < 0.05$, ** with $p < 0.01$)	154
39.	Evaluation of semantic abstraction when trained on New York City and San Francisco data sets. (* significant difference in error rates with $p < 0.05$, ** with $p < 0.01$)	156
41.	Evaluation of semantic abstraction when trained on Seattle data sets. (* significant difference in error rates with $p < 0.05$, ** with $p < 0.01$)	157
43.	Number of tokens both data sets have in common.	159
44.	Number of tweets containing location mentions and temporal expressions as well as LOD types and categories.	160
45.	Number of distinct types and categories extracted for both data sets.	160

46.	The most representative LOD features for incident-related and not incident-related tweets in each data set.	161
47.	F-measures for training and testing on one data set using 10f-CV. . . .	162
48.	F-measures for training on one city and testing on a different city. . . .	162
49.	F-measures for training on one city and testing on a different city after manual feature selection of LOD features.	165
50.	Overview of related approaches with respect to the type of event, clustering approaches, and metadata used.	173
51.	Overview of related approaches with respect to the approaches used for event detection and clustering.	174
52.	Content characteristics of tweets differentiated by user type.	197
53.	Content characteristics of incident reports and tweets not related to incidents.	198
54.	Percentage of situational feature categories for each user type per category and in relation to the overall amount of tweets per user type (in parentheses).	200
55.	Correlation results of real-world incidents to incidents mentioned in tweets.	205
56.	Comparison of approaches for small-scale incident detection.	206
57.	Comparison of related active learning approaches with respect to selection strategies and the use of event-related metadata.	214
58.	Overview of the error rates with changing number of annotators (10 repetitions).	228
59.	Deficiencies $DEF(AL)$ of related strategies and our approach. The approach of Tang et al. is used as a baseline strategy.	228
60.	Deficiency of our approach with the number of annotators (number of errors in parentheses). No noise is used as a baseline strategy. . . .	232
61.	Classification results for all features of plain supervised approach that uses all instances (training on 1,200 instances, test on 800 instances). . . .	235
62.	Comparison of using semantic abstraction compared to a baseline comprising n-gram features after Slang and URL replacement, TF-IDF scores, and syntactic features on 2-CLASSES data set with word-2-grams and SVM as classifier.	243
63.	Comparison of using semantic abstraction compared to a baseline comprising n-gram features after Slang and URL replacement, TF-IDF scores, and syntactic features on 2-CLASSES data set with word-3-grams and SVM as classifier.	244

64.	Comparison of using semantic abstraction compared to a baseline comprising n-gram features after Slang and URL replacement, TF-IDF scores, and syntactic features on 2-CLASSES data set with char-5-grams and NB as classifier.	244
65.	Comparison of using semantic abstraction compared to a baseline comprising n-gram features after Slang and URL replacement, TF-IDF scores, and syntactic features on 4-CLASSES data set with word-2-grams and SVM as classifier.	245
66.	Comparison of using semantic abstraction compared to a baseline comprising n-gram features after Slang and URL replacement, TF-IDF scores, and syntactic features on 4-CLASSES data set with word-3-grams and SVM as classifier.	245
67.	Comparison of using semantic abstraction compared to a baseline comprising n-gram features after Slang and URL replacement, TF-IDF scores, and syntactic features on 4-CLASSES data set with char-5-grams and NB as classifier.	246
68.	Evaluation results of machine-based aggregation using leader-follower clustering.	247
69.	Evaluation results of machine-based aggregation using DBScan clustering.	248

List of Algorithms

1. Algorithm for rule-based clustering. 187
2. Algorithm for initial selection strategy. 221
3. Algorithm for one iteration of the query selection strategy. 223



Curriculum vitae⁵⁹

- 2004-2011** Technische Universität Darmstadt
Diploma, Business and Computer Science (Dipl. Wirtsch.-Inform.)
- 2011-2014** SAP AG, Darmstadt
Research Associate and Ph.D. Student
- 2011-2014** Technische Universität Darmstadt
Ph.D. Student, Telecooperation Lab

⁵⁹ Gemäß §20 Abs. 3 der Promotionsordnung der TU Darmstadt